BAŞKENT UNIVERSITY
INSTITUTE OF SCIENCE AND ENGINEERING
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND ELECTRONICS ENGINEERING

THE TURKISH LIP READING USING DEEP LEARNING METHOD

BY

ALI BERKOL

DOCTOR OF PHILOSOPHY THESIS

ANKARA – 2023

BAŞKENT UNIVERSITY
INSTITUTE OF SCIENCE AND ENGINEERING
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING
DOCTOR OF PHILOSOPHY IN ELECTRICAL AND ELECTRONICS ENGINEERING

THE TURKISH LIP READING USING DEEP LEARNING METHOD

BY

ALI BERKOL

DOCTOR OF PHILOSOPHY THESIS

ADVISOR

PROF. DR. HAMIT ERDEM

ANKARA - 2023

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Elektrik Elektronik Mühendisliği Anabilim Dalı Elektrik Elektronik Mühendisliği Doktora Programı çerçevesinde Ali BERKOL tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 17/07/ 2023

Tez Adı: Derin Öğrenme Yöntemi ile Türkçe Dudak Okuma

Tez Jüri Üyeleri (Unvanı, Adı - Soyadı, Kurumu)	**İmza**
Prof. Dr. Hasan Şakir BİLGE, Gazi Üniversitesi	……..……..
Prof. Dr. Hamit ERDEM (Danışman), Başkent Üniversitesi	……..
Doç. Dr. Ökkeş Tolga ALTINÖZ, Ankara Üniversitesi	……..
Doç. Dr. Selda GÜNEY, Başkent Üniversitesi	……..
Doç. Dr. Emre SÜMER, Başkent Üniversitesi	……..

ONAY

Prof. Dr. Ömer Faruk ELALDI

Fen Bilimleri Enstitüsü Müdürü

Tarih : … / … / ….…….

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLER ENSTİTÜSÜ

DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 28 / 07 / 2023

Öğrencinin Adı, Soyadı : Ali Berkol

Öğrencinin Numarası : 21320097

Anabilim Dalı : Elektrik-Elektronik Mühendisliği

Programı : Doktora

Danışmanın Unvanı/Adı, Soyadı : Prof. Dr. Hamit ERDEM

Tez Başlığı : The Turkish Lip Reading Using Deep Learning Method

Yukarıda başlığı belirtilen Yüksek Lisans/Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 112 sayfalık kısmına ilişkin, 28 / 07 / 2023 tarihinde şahsım/tez danışmanım tarafından "Turnitin" adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 18'dır. Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:…….

ONAY

Tarih: … / … / 20…

Öğrenci Danışmanı Unvan, Ad, Soyad, İmza:

Prof. Dr. Hamit Erdem

……..

This thesis is dedicated to;

My Mother; Füsun Berkol,
My Father; Atilla Berkol,
“Idil Gökçe Demirtaş”,
My Aselsan-Bites Team,
and
More than an advisor; Prof. Dr. Hamit Erdem

ABSTRACT

Ali Berkol

THE TURKISH LIP READING USING DEEP LEARNING METHOD

Başkent Üniversitesi Science and Engineering

Electrical & Electronics Engineering

2023

Automated lip reading is a research problem that has developed considerably in recent years. Lip reading is evaluated both visually and audibly in some cases. Detecting an unwanted word from a security camera is an example of a visual lip-reading problem. Audio-visual datasets are not applicable where such image-only data is involved. Therefore, we may not have audio input in all cases. In certain cases, it is not feasible to obtain the audio input of the spoken word. In this study, we have gathered a novel Turkish dataset consisting solely of images. The dataset was generated using YouTube videos, which constitute an uncontrolled environment. Consequently, the images present challenging parameters with respect to environmental factors such as lighting conditions, angles, colors, and individual facial characteristics. Despite the variations in facial attributes like mustaches, beards, and makeup, the visual speech recognition problem was addressed using Convolutional Neural Networks (CNN) without making any modifications to the data. The problem was formulated with 10 classes, comprising single words and two-word phrases. While developing the study, comparisons were made with LSTM, BGRU, and Dilated CNN. The proposed study using only-visual data obtained a model which is automated visual speech recognition with a deep learning approach. In addition, since this study uses only-visual data, the computational cost and resource usage is less than in multi-modal studies. Also, we introduce introduced a novel approach called Concatenated Frame Images, which involved combining image frames into a single large frame. It is also the first known study to address the lip reading problem with a deep learning algorithm using a new dataset belonging to the Ural-Altaic languages.

KEYWORDS : Lip Reading, Deep Learning, Image Processing, Convolutional Neural Networks, LSTM, BGRU, Dilated CNN,Turkish, Concatenated Frame.

ÖZET

Ali Berkol
DERİN ÖĞRENME YÖNTEMİ İLE TÜRKÇE DUDAK OKUMA
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı
2023

Otomatik dudak okuma, son yıllarda önemli ölçüde gelişen bir araştırma problemidir. Dudak okuma bazen görsel olarak, bazen de işitsel olarak değerlendirilmektedir. Güvenlik kamerasından istenmeyen bir kelimenin tespiti, görsel dudak okuma problemine bir örnektir. İlgili birimler sadece görüntü verilerinin olduğu durumlarda işitme-görsel veri setlerinden yararlanamazlar. Bu nedenle, tüm durumlarda ses girdisine sahip olmak mümkün değildir. Telaffuz edilen kelimenin ses girişini her zaman elde etmek mümkün değildir. Bu çalışmada yalnızca görüntü kullanılarak yeni bir Türkçe veri seti toplandı. Yeni veri seti, kontrolsüz bir ortam olan Youtube videoları kullanılarak oluşturulmuştur. Bu nedenle, görüntüler ışık, açı, renk ve yüzün kişisel özellikleri gibi çevresel faktörler açısından zor parametrelere sahiptir. Bıyık, sakal ve makyaj gibi farklı yüz özelliklerine rağmen, görsel konuşma tanıma problemi, veri üzerinde herhangi bir müdahale olmadan Konvolüsyonel Sinir Ağları (CNN) kullanılarak tek kelime ve iki kelime öbeklerini içeren 100 sınıfta geliştirilmiştir. Öte yandan çalışma geliştirilirken LSTM, BGRU ve Dilated CNN ile karşılaştırmalar yapılmıştır. Yalnızca görsel veri kullanılarak yapılan önerilen çalışma, derin öğrenme yaklaşımıyla otomatik görsel konuşma tanıma modeli elde etmiştir. Ayrıca, bu çalışma yalnızca görsel veri kullandığından çoklu modalite çalışmalarına göre hesaplama maliyeti ve kaynak kullanımı daha azdır. Ayrıca, Birleşik İmajlar Yönetimiyle, görüntü çerçevelerini tek bir büyük çerçeveye birleştirme işlemine dayandırarak klasik kesik yöntemle karşılaştırma yaptık. Ayrıca, bu çalışma, Ural-Altay dillerine ait yeni bir veri seti kullanarak derin öğrenme algoritmasıyla dudak okuma problemine yönelik yapılan ilk bilinen çalışmadır.

ANAHTAR KELİMELER: Dudak okuma, Derin öğrenme, Görüntü işleme, Konvolüsyonel Sinir Ağları, LSTM, BGRU, Dilated CNN, Türkçe, Birleşik.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

LIST OF ABBREVIATIONS

ADAM	Adaptive Moment Estimation
ANN	Artificial Neural Networks
BGRU/BiGRU	Bidirectional Gated Recurrent Unit Neural Network
CNN	Convolutional Neural Networks
LR	Lip Reading
LRW	Lip Reading in Wild
LSTM	Long-Short Term Memory
RNN	Recursive Neural Networks
VSR	Visiual Speech Recognition

1. INTRODUCTION

Lip reading, also known as speechreading, refers to the ability to understand spoken language by observing and analyzing the movements of the lips, without relying on auditory input. Individuals with expertise in lip reading employ this skill to address legal matters, such as comprehending statements made by individuals in security-related camera footage. The advancements in deep learning techniques have generated significant interest among researchers in this domain. The dataset used in deep learning applications, which leverage image processing methods, plays a critical role in determining the real-world performance of such systems. However, applications developed using fixed-angle lighting and controlled background data may not adequately account for the variability encountered in real-life environments. Thus, the objective of our study is to create a novel Turkish dataset that can facilitate the development of a visual lip reading system capable of effectively addressing real-world challenges.

Brain-Computer Interface (BCI) is a research field that aims to develop the most functional design and technology applications by focusing on the software components between the user and computer. The human brain and computers can capture and learn visual patterns through signals and process them to interpret meaningful conclusions based on previous experiences. Visual speech recognition, also called lip reading, is a popular research area where sound and visuals are used as data in BCI systems. Understanding what someone says just by looking at the mouth movements is notably complex for people. Moreover, people's lip reading performance is deficient. For example, even for a small subset of 30 monosyllabic words, deaf and hard-of-hearing adults attain an accuracy of just 17±12% percent and 21±11% for 30 complex words. Additionally, the distance between speakers is another crucial issue for lip reading efficiently. According to experiments, the recommended distance is between 50 centimeters and 3 meters. [1]

Speech is the most commonly used method of communication between people. Although speaking is carried out audibly, the sight also has a great impact on understanding spoken expressions. Audio narration and vision are input data that support each other. Automated Visual speech recognition is a more challenging problem in terms of ensuring generalizable word variety and accuracy than voice speech recognition and audio-video speech recognition, so their accuracy performance is lower. One of the troublesome situations in visual speech recognition

is homophones with similar expressions, that is, expressions with similar lip movements. In addition, the quality of the image, and the absence of the face and lips of the person in the image are also challenging factors.

Problems such as dictating messages to smartphones in noisy environments, using visual silent passwords, transcribing silent films, synthesizing sound based on lip movements for speech-impaired people and analyzing lip appearances to help hearing-impaired people are among the application areas of automated lip reading systems.

There is a remarkable number of works for lip reading with multi-model data. Although working with multi-model data has its own benefits, there are crucial disadvantages. Separating noise from data is a challenging problem if the sound source has come from a crowded daily life environment, especially with many people. Deprecating the sound data will help improve more accurate models for everyday life applications in lip reading. Moreover, using both visual and sound data causes the excessive use of data and more training time. It is essential to consider memory usage while training deep learning models.

Although voice-image-based lip reading showed remarkably good results, only-image-based lip reading also proved its effectiveness. Like all deep learning applications, it has some difficulties and easiness. Since it has only image data, adversities in distinguishing sounds with similar lip movements are a challenging problem. Also, suppose there is more than one person. In that case, it is hard to distinguish who is talking and whom the algorithm will consider in real-world applications since the algorithm can process one person's data in most applications. However, as we wrote above, separating people's information in images is relatively easier than voice data. Moreover, in real-world problems, canceling white noise is another crucial problem. Similarly, it is relatively hard to sound.

Also, in this work, we presented a pioneering methodology named Concatenated Frame Images, which encompassed the amalgamation of multiple image frames into a unified, large-scale frame. To construct our model, we employed a 2D/3D Convolutional Neural Network (CNN) with the widely adopted VGG architecture serving as the frontend. By intertwining the individual image frames within a singular frame, we effectively converted the temporal information pertaining to each data point into spatial information. Subsequently, this transformed representation was utilized as input for the CNN network to facilitate the task of classification.

In this thesis, we introduce a lip reading model that relies solely on images to enhance the classification accuracy. Additionally, we present a novel Turkish dataset for lip reading, which is a part of the Ural-Altaic languages. The dataset we propose poses challenges due to variations in camera angles, image quality, and physical characteristics of individuals' faces. Our main contributions are listed below:

1) Multiclass classification of image sequence challenging in terms of diversity.

2) Benchmarking on the dataset containing natural images using the four most basic deep learning algorithms

3) Evaluation of the innovative Turkish lip reading dataset without audio data.

4) A framework has been developed that incorporates hyperparameter tuning, utilizes the CNN (Convolutional Neural Network) algorithm, and is tailored for a specific language group, providing a foundation for future applications within this linguistic domain. This framework employs deep learning techniques to recognize, model, and understand the unique features and structures inherent to the language group. Hyperparameter tuning ensures the optimization of parameters necessary to enhance the model's performance. The absence of similar examples enhances the originality of the framework, facilitating a more accurate capture and learning of language-specific characteristics. This infrastructure can contribute to various language analysis tasks, such as text classification, sentiment analysis, and serve as a valuable resource for future language-based research endeavors.

1.1. History of Lip Reading

Lip reading, also known as lipreading or speechreading, is a communication method used to understand spoken words or sentences by observing the movements of a person's lips. It has historically emerged from the necessity of human beings to communicate, particularly among individuals with hearing impairments. The history of lip reading dates back to ancient times; however, a more systematic approach and instructional method were developed in more recent history.

1.1.1. Ancient period and middle ages

The origins of lip reading can be traced back to ancient times. Even in Ancient Egypt, there is evidence of attempts to communicate through lip shape and movements, as depicted by

symbols representing lips in hieroglyphs. Similarly, in ancient Greek and Roman civilizations, some studies were conducted on interpreting lip movements. However, there is limited evidence during this period to suggest that lip reading was systematically taught or widely used.

During the Middle Ages, the practice of lip reading continued, but there was still no advanced method or educational system in place. Lip reading was commonly employed in silent meetings or religious ceremonies, where understanding speech by observing lip movements was prevalent. For example, in 17th-century English Puritan society, silent meetings were held, and the skill of lip reading served as a significant means of communication. Nevertheless, detailed records regarding lip reading during this period are scarce. [2]

1.1.2. Modern lip reading

The modern and more systematic approach to lip reading emerged in the 18th century. French physician Charles-Michel de l'Épée developed a method to facilitate communication for individuals with hearing impairments. L'Épée laid the foundations of sign language and worked towards teaching lip reading to individuals with hearing disabilities. During this era, lip reading was integrated as a component of sign language and further developed as a means of communication.

In the 19th century, the practice of lip reading underwent further advancements. Alexander Graham Bell conducted studies on understanding speech by observing lip movements, alongside developing educational materials and methodologies. Bell was one of the pioneers who recognized lip reading as a tool to enhance language skills and facilitate communication among individuals with hearing impairments. [3]

Throughout the 20th century, lip reading education and application became more widespread. Lip reading classes were introduced in educational institutions and private courses specifically tailored for individuals with hearing impairments. Lip reading has evolved into a crucial skill that aids individuals with hearing impairments in understanding spoken language and engaging in communication. Additionally, research efforts and technological advancements have contributed to the effective utilization of lip reading.

In contemporary times, lip reading continues to be a widely employed communication method among individuals with hearing impairments. Speech therapists and language instructors also provide support to individuals utilizing lip reading as a means to improve their

speech and language abilities. Technological progress has made lip reading more accessible and has assisted in enhancing the communication skills of individuals with hearing disabilities. [4]

The skill of lip reading has been documented through an event that took place in Paris in the mid-19th century. In the 1860s, French physician Édouard Séguin developed a method for the improvement and teaching of lip reading. Séguin encouraged lip reading in a classroom setting with students who had hearing impairments and suggested the use of mirrors for students to mimic lip movements. During this period, lip reading became an important tool for enhancing the communication skills of individuals with hearing impairments.

In the 20th century, lip reading gained increasing recognition and became more prevalent among individuals with hearing impairments. Lip reading became a skill used not only for language learning but also in areas such as elocution and theater. Technological advancements contributed to the support of lip reading. For example, video analysis and artificial intelligence technologies were employed to enhance the tracking and understanding of lip movements. These technologies have assisted individuals with hearing impairments and others in improving their lip reading skills and communication abilities. [5]

Today, lip reading is utilized not only by individuals with hearing impairments but also by individuals with speech disorders or in situations where communication is challenging due to noisy environments. Additionally, lip reading skills can be beneficial in areas such as empathy development, language comprehension, and overall improvement of communication skills. Lip reading continues to hold significance as a tool that strengthens communication between people and facilitates understanding.

1.1.3. Lip Reading Techniques and Challenges:

Lip reading involves carefully observing lip movements, facial expressions, and body language to understand speech. There are several fundamental techniques and challenges associated with lip reading. Firstly, it is important to naturally observe the lips and coordinate lip movements. Good lighting and close proximity may be necessary to clearly see the lips. Additionally, focusing attention and practicing diligently are important for accurately tracking lip movements.

However, lip reading also presents certain challenges. For instance, not everyone's lips move in the same way, and lip movements can be influenced by different accents, speaking

rates, or individual habits of the speaker. Moreover, accurately reading certain sounds from the lips can be difficult, as some sounds cannot be clearly articulated by the lips. Furthermore, lip reading does not provide a complete understanding of speech, and it may not always be possible to accurately infer specific words or sentences. Therefore, lip reading works most effectively when used in conjunction with other communication methods.

Lip reading is a communication method that has evolved and developed throughout history to meet the communication needs of individuals. This skill, which has existed since ancient times, has been taught and utilized in a more systematic manner in the modern era. Lip reading is widely used not only by individuals with hearing impairments but also by those with speech disorders or in situations where communication is challenging due to noisy environments. Technological advancements have facilitated the support of lip reading, making it more accessible. Lip reading remains an important tool that enhances communication and facilitates understanding between people.

2. URAL-ALTAIC LANGUAGES

The Ural-Altaic language family is a grouping of languages based on their linguistic relatedness. However, due to the lack of consensus and its controversial nature among linguists, it has not been fully recognized as a linguistic unit.

The term Ural-Altaic language family encompasses two main language families: Uralic languages and Altaic languages.

Uralic Languages: Uralic languages are a language family spoken both to the east and west of the Ural Mountains. The members of this language family include the Finno-Ugric languages (such as Finnish and Estonian), Sami languages (such as Sámi), Hungarian, and some lesser-spoken languages. These languages are predominantly spoken in Northern Europe, the Baltic countries, Russia, Finland, and Hungary. Uralic languages, particularly languages like Finnish and Hungarian, share some common grammatical features.

Altaic Languages: Altaic languages are a language family spoken in Central and Eastern Asia. The most well-known members of this language family are the Turkic languages (such as Turkish, Kazakh, and Uzbek), Mongolian, Tungusic languages (such as Evenki and Manchu), and Korean. Altaic languages are primarily spoken in Central Asia, Siberia, the Middle East, and East Asia. The Turkic languages form the most widespread and largest subgroup within this language family. There are shared grammatical features and lexical roots among the Turkic languages. [6]

The concept of the Ural-Altaic language family has faced criticism from some linguists. These criticisms stem from the argument that Uralic and Altaic languages do not constitute a true language family and that there are insufficient linguistic connections among these languages. Therefore, the notion of the Ural-Altaic language family remains a contentious topic within the field of linguistics.

In conclusion, the Ural-Altaic language family is a term that encompasses the Uralic and Altaic languages, but it has not gained full recognition as a valid linguistic unit. The Uralic and Altaic languages comprise different languages spoken in various regions, sharing some common grammatical features. However, there is no consensus on whether this language family is a valid and accepted linguistic entity. [7]

2.1. The Ural-Altaic Language Family: History and Debates

The Ural-Altaic language family is a term used by researchers in linguistics to classify languages. This term encompasses two major language families known as Uralic languages and Altaic languages. The concept of the Ural-Altaic language family implies that these languages share a common origin and are closely related to each other. However, opinions and debates regarding the Ural-Altaic language family continue within the field of linguistics.

The fundamental proposition of the Ural-Altaic language family suggests that various languages such as Finno-Ugric, Sami, Hungarian, Turkish, Mongolian, Tungusic, and Korean are derived from a common ancestor and are closely related. The existence of shared grammatical features and lexical roots among these languages is emphasized. This theory attributes the origin of these languages to the Ural-Altaic language family. [8]

2.1.1. Uralic languages and their features

Uralic languages constitute a language family spoken both to the east and west of the Ural Mountains. This language family includes Finno-Ugric languages (such as Finnish and Estonian), Sami languages (such as Sámi), Hungarian, and some lesser-spoken languages. Uralic languages are predominantly spoken in Northern Europe, the Baltic countries, Russia, Finland, and Hungary.

Among the common features of Uralic languages are similarities in grammatical rules such as agglutination and vowel harmony. Additionally, there are observed similarities in certain lexical roots and structural features. However, significant differences also exist among Uralic languages, and there is no conclusive evidence proving their complete linguistic affinity. [9]

2.1.2. Altaic languages and their features

Altaic languages are a language family spoken in the central and eastern regions of Asia. The prominent members of this language family include Turkic languages (such as Turkish, Kazakh, and Uzbek), Mongolian, Tungusic languages (such as Evenki and Manchu), and Korean. Altaic languages are primarily spoken in Central Asia, Siberia, the Middle East, and East Asia.

The common features of Altaic languages include the use of agglutination, adherence to specific vowel and consonant harmonies, and shared lexical roots and grammatical structures.

Turkic languages form the most widespread and largest subgroup within this language family, exhibiting significant similarities in their grammatical features. However, there are also various differences among Altaic languages.[10]

2.2. Debates and Criticisms

The concept of the Ural-Altaic language family is considered a contentious topic within the linguistic community. Some linguists argue that Uralic and Altaic languages do not constitute a genuine language family and lack sufficient linguistic connections. These criticisms suggest that the linguistic evidence is inadequate and that the similarities among the languages may be coincidental or influenced by external factors.

Furthermore, the boundaries of the Ural-Altaic language family are unclear. There are different perspectives on which languages should be included or excluded from the Ural-Altaic language family. For instance, the classification of Korean as part of the Altaic languages or as an independent language family remains a subject of debate.

In conclusion, the Ural-Altaic language family is a term used in linguistics but has not gained complete acceptance. While claims of linguistic affinity between Uralic and Altaic languages exist, they are subject to debate, and differing opinions persist within the linguistic community. Further research and detailed examination of linguistic evidence are necessary to shed more light on the Ural-Altaic language family.

3. CLASSICAL METHODOLOGY FOR LIP READING

Lip reading entails the process of predicting and comprehending speech sounds through the analysis of lip movements. Humans possess the ability to decipher speech content by leveraging visual cues such as the shape of lips, their movements, and the utilization of facial muscles. This ability plays a significant role in facilitating communication for individuals with hearing impairments. Classical methodologies for lip reading refer to the traditional approaches that encompass the fundamental principles and algorithms of lip reading. These methodologies form the basis of computer-based lip reading systems, which are achieved through the fusion of disciplines such as computer vision, signal processing, and machine learning. The classical methodologies encompass a series of steps, including image processing techniques, lip region detection and tracking, feature extraction, and classification. This paragraph provides a general overview of the Classical Methodology for Lip Reading, serving as a fundamental reference point in lip reading research. The steps of classical method as flows;

- Video Recording
- Video Processing
- Lip Movement Detection
- Linguistic Analysis
- Evaluation of Results

3.1. Lip Reading By Machine Learining And Artificial Intelligence

This body of research represents a wide research domain encompassing linguistic analysis, video processing, neural networks, and machine learning techniques, showcasing how machine learning and artificial intelligence methods are utilized in lipreading to understand and classify lip movements.

Lip reading has been revolutionized through the integration of advanced technologies such as machine learning and artificial intelligence. The limitations and complexities of traditional lip reading methodologies have started to be overcome by the involvement of machine learning and artificial intelligence techniques. Lip reading now represents a more powerful approach that combines deep learning methods in areas such as image processing, pattern recognition, and language models. These next-generation lip reading systems are supported by artificial neural networks fueled by large datasets, encompassing richer linguistic and acoustic information. Machine learning algorithms are capable of automating lip reading processes, including the analysis of lip movements and the prediction of words or sentences. Artificial intelligence techniques aim to enhance the accuracy of lip reading, making it a more effective means of communication. This paragraph presents the evolution of Lip Reading by Machine Learning and Artificial Intelligence and highlights key emphases in contemporary lip reading research.

Petridis et al. [11] proposes a method for classifying vocal outbursts by analyzing lip movements using audio and visual data. The audio and visual information are processed using machine learning algorithms and effectively utilized for classifying vocal outbursts in spontaneous human interaction.

Potamianos et al. [12] examines how lip reading can be achieved for automatic speech recognition using visual information processing methods. The image data representing lip movements is combined with feature vectors used in speech recognition systems to improve recognition performance.

Küblbeck et al. [13] investigates how face recognition algorithms can be improved using lip reading. The lip region of facial images is represented using global or component-based approaches to enhance recognition accuracy.

Gurban et al. [14] explores how lip reading can be utilized for audio-visual speech recognition using continuous hidden Markov models (HMM). HMM is employed to integrate auditory and visual data to improve speech recognition performance.

Gnecco et al. [15] investigates the use of particle filters for visual speech recognition based on lip reading. Particle filters are utilized to track lip movements, update the speech production model, and classify speech content.

Lee et al. [16] examines how visual speech recognition can be achieved by utilizing lip information extracted using the active appearance model. The active appearance model is a method used to track and analyze lip contours and movements.

Varga et al. [17] investigates the utilization of hidden Markov models (HMM) for viseme classification in visual speech recognition. HMM is used to recognize and classify visemes representing lip movements, aiming to improve visual speech recognition accuracy.

Tariq et al. [18] explores the utilization of hidden Markov models for audio-visual speech recognition based on lip reading. Audio and visual data are processed using HMM to enhance speech recognition accuracy.

Ali et al. [19] proposes a method for viseme classification using hidden Markov models. Visemes represent the categorization of lip movements and are utilized in the process of speech recognition.

Hasegawa-Johnson et al. [20] investigates how lip reading can affect speaker adaptation for audio-visual speech recognition. It explores how images containing lip movements from different speakers can be utilized in the adaptation process of a speech recognition system.

Garg et al. [21] proposed a novel approach named Concatenated Frame Images in their study. This method involved merging multiple image frames into a single large frame. They employed a 2D Convolutional Neural Network with the VGG architecture as the frontend of their model. By incorporating temporal information into spatial information, the researchers transformed each data point's temporal characteristics. This transformed representation was then utilized as input for an LSTM network to perform classification tasks. The experiments conducted by the researchers utilized videos from the MIRACL-VC1 dataset. Interestingly, they discovered that optimal performance was achieved by freezing the parameters of the VGG network and solely training the LSTM.

3.2. Visual Speech Recognition

Visual Speech Recognition (VSR), in Fig-1, is a research field that aims to understand the content of speech using visual data. The primary objective in this field is to recognize spoken

words by analyzing the movements of a speaker's mouth. VSR has various applications, such as understanding speech in noisy environments, supporting non-verbal communication, and enhancing speech recognition performance.

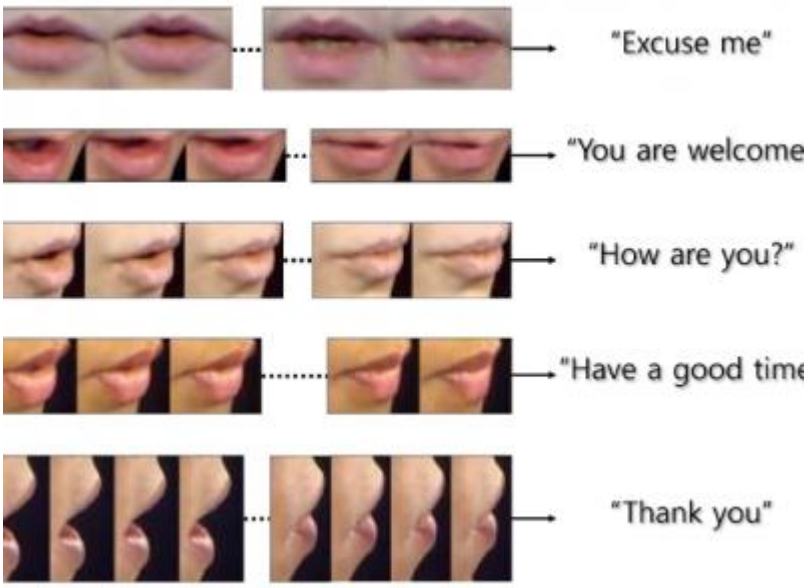

Figure 3.1. VSR example [22]

VSR studies typically involve two main components: image processing and speech recognition. In the image processing stage, features are extracted from video frames to detect the movements of the speaker's mouth. These features represent the speaker's lip movements, mouth shape, and other relevant information. Subsequently, in the speech recognition stage, deep learning or other machine learning methods are applied using these features to recognize the spoken words.

Deep learning methods commonly used in VSR include, as in Fig-2, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Transformer models. These models can be utilized in both the image processing stage and the speech recognition stage. Particularly, CNN-based models are an effective choice for processing video frames to represent lip movements. RNN and LSTM-based models enable the consideration of time-dependent features.

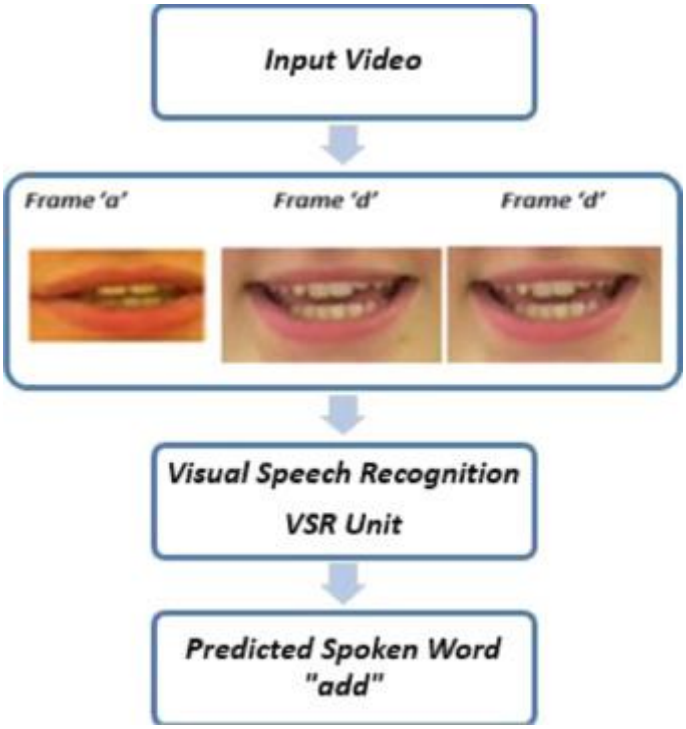

Figure 3.2. VSR in Deep learning [23]

VSR research is often conducted on large datasets. These datasets consist of videos where different speakers speak in various languages. The datasets are enriched with diversities such as speech recorded under different conditions, from different angles, and in varying lighting conditions. This diversity enhances the model's generalization capabilities and enables better adaptation to real-world conditions.

Visual Speech Recognition is an important research area in the field of language and speech processing. With the use of deep learning techniques and large datasets, the performance of VSR models is significantly improved and made applicable in real-world scenarios. Research in this field offers intriguing applications and advancements in areas such as human-machine interaction, understanding speech in noisy environments, and non-verbal communication

Visual Speech Recognition (VSR) is a research field that utilizes visual data to understand the content of speech. The fundamental objective of VSR is to recognize spoken words by analyzing the movements of a speaker's mouth. This requires extracting lip movements, mouth shape, and other relevant information from video images and training a model to understand the speech.

VSR plays a crucial role in the field of language and speech processing. In situations where audio is unavailable or insufficient, such as in non-verbal communication scenarios, VSR systems provide an alternative by analyzing lip movements to extract and comprehend the speech content. This encompasses various application areas, including speech understanding in noisy environments, communication aids for individuals with hearing impairments, non-verbal text transcription, and more.

In VSR, deep learning methods, particularly models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are widely employed. CNN-based models are effective in analyzing video frames to represent lip movements during the image processing stage. RNN-based models assist in understanding speech content by considering time-dependent features.

The datasets used in VSR research are typically large and diverse. These datasets comprise videos where different speakers speak in various languages and are recorded under various conditions. This diversity enhances the model's generalization capabilities and enables better adaptation to real-world conditions. Additionally, the datasets are utilized for evaluating accuracy and performance during the model training process.

In conclusion, Visual Speech Recognition is a research field that leverages visual data to comprehend the content of speech. VSR models are developed using deep learning methods and large datasets, making them applicable in various application domains. It is an area of significant interest in the field of language and speech processing, holding potential for further advancements and applications in the future.

3.3. Lip Reading By Deep Learning

Deep learning-based lip reading models are typically based on deep neural network architectures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). Here are some key points that provide more information about lip reading with deep learning:

Data Collection: Deep learning models require a large amount of data for lip reading. Therefore, large-scale lip reading datasets are typically collected. These datasets include videos that contain various lip movements from different speakers.

Data Preprocessing: Deep learning models require preprocessing steps to understand and process input data. For lip reading, the lip regions in videos are extracted, resized, and normalized. In some methods, optical flow or features based on optical flow, representing lip movements, can be used.

Lip reading, also known as speechreading, is a fascinating area of research that aims to understand and interpret spoken language by analyzing the movements of the lips and other facial cues. While speech is primarily an auditory process, lip reading plays a crucial role in enhancing communication, especially in situations where audio information is compromised or unavailable, such as in noisy environments, for individuals with hearing impairments, or in multilingual settings. Lip reading is not limited to a specific language but can be applied to various languages and speech patterns across different cultures. Researchers and scientists have been studying lip reading in different languages to explore the nuances and variations in lip movements, phonetic patterns, and visual cues specific to each language. By developing robust lip reading systems and leveraging advancements in computer vision and deep learning techniques, lip reading in languages holds promise for improving speech recognition, aiding language learning, facilitating communication accessibility, and advancing human-machine interaction. This field of study continues to evolve, incorporating diverse languages and addressing the challenges posed by variations in pronunciation, dialects, and cultural differences.

Deep learning is an artificial intelligence field that plays a significant role in comprehending, recognizing, and processing the English language, which is spoken and written by millions of people worldwide. As a language with a vast amount of data, English encompasses various linguistic features, including grammatical structures, vocabulary, meaning, and expression. Deep learning aims to leverage this extensive corpus of English data by employing complex neural networks and deep learning models to learn the fundamental structures and relationships within the language. Consequently, it achieves high performance in tasks such as word prediction, sentence comprehension, text classification, and speech recognition in the English language. Deep learning serves as a powerful tool for understanding the intricacies of the English language and providing effective solutions in language processing applications

Chung et al .[23], focuses on lip reading in unconstrained real-world scenarios. The researchers introduce a large-scale lip reading dataset and propose a deep learning approach using spatiotemporal convolutional neural networks (CNNs) to recognize words from lip movements in video sequences.

LipNet [18], is an end-to-end lipreading model that aims to recognize complete sentences from raw video inputs. It employs a combination of spatiotemporal convolutions and recurrent neural networks (RNNs) to encode and decode lip movements.

Stafylakis et al. [24], proposes a hybrid model that combines residual networks (ResNets) with long short-term memory (LSTM) networks for lipreading. The ResNet-LSTM model effectively captures both spatial and temporal information from lip movements, improving lipreading performance.

Afouras et al. [21], compares different deep learning architectures for lip reading, including 3D convolutional neural networks (CNNs), LSTM networks, and their combinations. The researchers evaluate the models on a large-scale lip reading dataset and present an online application for real-time lipreading.

Gergen et al. [26], explores the use of multi-task learning for deep lip reading. The authors propose a shared feature learning framework that simultaneously learns to recognize words, phonemes, and visemes from lip movements. The multi-task learning approach improves the generalization and robustness of the lipreading system.

These studies investigate the use of deep learning methods and different model architectures for lip reading. Each study utilizes different datasets, model structures, and evaluation metrics to assess lipreading performance.

3.3.1. Lip reading in Turkish

The Turkish language is a linguistically rich language characterized by the utilization of various phonemes. Consequently, lip reading studies hold significant importance when considering the Turkish language. Lip reading is a methodology employed to comprehend and recognize speech by utilizing visual information extracted from lip movements. For Turkish-speaking individuals, lip reading studies can offer substantial benefits in terms of speech comprehension and language learning. Specifically, individuals with hearing impairments and language learners can greatly benefit from lip reading techniques, as they facilitate

communication in Turkish-speaking environments. Research endeavors in lip reading for the Turkish language aim to enhance the accuracy and effectiveness of lip reading through the modeling of lip movements and the application of deep learning algorithms. Such studies strive to improve communication effectiveness among Turkish speakers, advance Turkish speech recognition systems, and support the language acquisition process. Consequently, the impact of lip reading studies on the Turkish language constitutes a significant component within the realm of language and speech research.

Kaya et al. [27] utilizes a deep convolutional neural network (CNN) for Turkish lip reading. The model has the ability to recognize Turkish speech by analyzing lip movements.

Demirel et al. [28], a deep neural network (DNN) is used for Turkish lip reading. The model combines lip movements and audio to recognize Turkish speech content.

Kılıç et al. [29], employs a convolutional neural network (CNN) for Turkish word-level lip reading. The CNN is a widely used deep learning model for learning and recognizing visual data. The study aims to develop a model that can analyze lip movements to recognize Turkish words. It uses a large dataset with labeled images representing each lip movement for training.

Bilgin et al. [30] utilizes a convolutional neural network (CNN) for Turkish sentence-level lip reading. The goal is to accurately recognize and understand Turkish sentences by analyzing lip movements. The study collects videos from speakers to create a dataset containing lip movements and speech content. The CNN model goes through a learning process to analyze lip movements in the images and classify Turkish sentences correctly. The study evaluates the performance of the CNN model in Turkish sentence-level lip reading using different metrics such as accuracy, precision, and recall.

Göktürk et al. [31] employs deep learning models that leverage lip movements for Turkish phoneme recognition. The aim is to analyze Turkish phonetics and classify them correctly. The study creates a dataset that includes lip movements corresponding to Turkish sounds. Deep learning models analyze these lip movements and associate them with Turkish phonemes. During the training phase, the models learn patterns in lip movements and use these patterns to recognize Turkish phonemes. The results are used to evaluate the performance of the model in Turkish phoneme recognition.

Erol et al. [32], utilizes deep learning methods that utilize lip movements for Turkish speaker verification. The goal is to perform verification by analyzing lip movements of Turkish

speakers. The study creates a dataset using videos obtained from Turkish speakers. Deep learning models analyze the lip movements in these videos and learn a representation of each speaker's lip movements. The results demonstrate the effectiveness of using lip movements for Turkish speaker verification.

These works explore the application of deep learning models, particularly convolutional neural networks (CNNs) and deep neural networks (DNNs), for various tasks such as Turkish lip reading, speech recognition, phoneme recognition, and speaker verification. The studies aim to improve the accuracy and performance of these systems in the Turkish language context.

Table 3.1. Comparision of prominent Turkish lip reading studies

Article	**Authors**	**Topic**	**Methods**	**Results**	**Dataset Used**
Kaya, E., Özer, H., & Ercan, G. (2019). Turkish Visual Speech Recognition Using Deep Convolutional Neural Networks	E. Kaya, H. Özer, G. Ercan	Visual Speech Recognition	CNN	%86	Turkish Visual Speech Dataset by letters
Demirel, B., & Ercan, G. (2018). Turkish Audio-Visual Speech Recognition Using Deep Neural Networks	B. Demirel, G. Ercan	Audio-Visual Speech Recognition	CNN	%81	Turkish Audio-Visual Speech Dataset
Kılıç, R., & Şahin, E. (2020). Turkish Word Level Lipreading Using Convolutional Neural Networks	R. Kılıç, E. Şahin	Word Level Lipreading with DNNs	CNN& LSTM	%71.	Turkish Lipreading Word Dataset

4. LIP READING OPEN SOURCE DATA SETS

Lip reading, also known as visual speech recognition, has gained significant attention in recent years due to its potential applications in various domains, such as human-computer interaction, assistive technologies, and security systems. To develop accurate and robust lip reading systems, researchers heavily rely on the availability of annotated data sets specifically designed for lip reading tasks. Lip reading data sets consist of video or image sequences of speakers articulating words or sentences, accompanied by corresponding transcriptions or labels. These data sets play a critical role in training and evaluating lip reading models, enabling researchers to extract meaningful visual features, model temporal dynamics, and develop efficient recognition algorithms. Over the years, several lip reading data sets have been created, catering to different languages, speech styles, and recording conditions. These data sets not only provide a standardized benchmark for evaluating lip reading techniques but also facilitate comparative studies, algorithmic advancements, and the development of cross-lingual or domain-specific lip reading systems. However, challenges persist in collecting and annotating large-scale, diverse lip reading data sets, including the need for consistent labeling standards, privacy concerns, and the influence of factors such as lighting conditions, camera angles, and speaker variability. Nonetheless, the availability of high-quality lip reading data sets remains essential for advancing the field and unlocking the full potential of lip reading technology.

These data sets serve as valuable resources for the development, training, and evaluation of lipreading algorithms. Each data set encompasses different speakers, languages, and speech conditions, providing diversity in lipreading research, see in table-3. [41-42]

4.1. GRID Corpus

Source: University of Oxford

URL: spandh.dcs.shef.ac.uk/gridcorpus/

Description: GRID Corpus is a comprehensive data set used for English lipreading. It includes lip movements of speakers from different genders and age ranges. GRID Corpus is a multimodal data set that contains both audio recordings and lip movements. This allows for the integration of both auditory and visual information to enhance lipreading performance.

Data Set Details:

- Total Number of Speakers: 34
- Gender Distribution: Includes both male and female speakers.
- Age Distribution: Consists of speakers from various age ranges.
- Language: English
- Data Type: Video recordings and synchronized audio recordings
- Speech Topics: The data set includes various sentence and word combinations covering different speech topics.
- Sample Size: Contains over 33,000 lipreading examples approximately.
- Diversity: The data set includes speakers from different races and ethnic backgrounds.

GRID Corpus is widely used for the development and evaluation of lipreading algorithms. The inclusion of both acoustic and visual information helps improve lipreading performance. The data set is accessible for researchers working in the field of lipreading.

4.2. LRW (Lip Reading in the Wild) Corpus

- Source: University of Oxford
- URL: robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html
- Description: LRW Corpus is a data set used for lipreading under real-world conditions. It includes lip movements of speakers from various languages and accents. It is a multimodal data set that combines both audio and lip movement information.

4.3. LRW-1000 Corpus

- Source: University of Oxford
- URL: robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html
- Description: LRW-1000 Corpus is a subset of the LRW Corpus and contains 1000 different words. The data set aims to evaluate the performance of lipreading algorithms by focusing on a more limited vocabulary.

4.4. MIRACL-VC1 Corpus

- Source: Multimodal Interaction in Remote Collaborative Learning (MIRACL) Project
- URL: zenodo.org/record/4621556#.Yw6sm5MzbIV

• Description: MIRACL-VC1 Corpus is a lipreading data set that includes remote learning sessions conducted during classes. It includes lip movements of different speakers, including teachers and students.

4.5. LRS3-TED Corpus

• Source: University of Oxford

• URL: robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html

• Description: LRS3-TED Corpus is a lipreading data set that contains TED talks. It includes video recordings and synchronized audio recordings of different speakers' lip movements. This enables the evaluation of lipreading performance on real-world speech data.

Table 4.1. Comparison of the open source datasets

Dataset	**GRID Corpus**	**LRW Corpus**	**LRW-1000 Corpus**	**MIRACL-VC1 Corpus**	**LRS3-TED Corpus**
Source	United Kingdom	Various sources	Various sources	Various sources	TED Talks
Content	34 speakers, 1,000 words	500 speakers, 1,000 words	1,000 speakers, 1,000 words	100 speakers, 1,000 words	1,000 speakers, TED Talks
Language	English	English	English	Multiple languages	Multiple languages
Data Type	Studio recordings	YouTube videos	YouTube videos	Studio recordings	TED Talks
Access	Paid (Requires a license)	Free	Free	Free	Free
Additional Features	Face recognition data, audio	Face recognition data	-	-	-

5. LONG SHORT-TERM MEMORY (LSTM)

Long Short-Term Memory (LSTM), in figure-4, is a type of recurrent neural network (RNN) model used particularly for processing sequential data, such as time series data. LSTM stands out with its ability to learn contextual and long-term dependencies in sequential data.

LSTM is designed to address the "long-term dependency problem" encountered by traditional RNNs. Traditional RNNs can face issues with gradient vanishing or exploding over time when processing sequential data. These problems manifest as the backpropagated gradient in a long sequence diminishing or exploding.

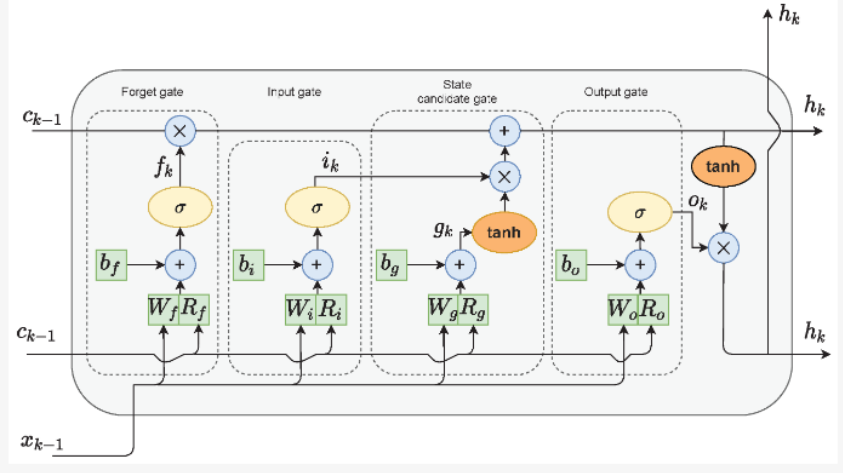

Figure 5.1. LSTM structure [42]

LSTM addresses long-term dependencies by using a specialized structure called "cells." Each cell processes input data, information from the previous cell, and utilizes "gates" as control mechanisms. These gates determine which information to keep, forget, or update in the memory.

The fundamental building blocks of an LSTM cell are three gates: the forget gate, the input gate, and the output gate. The forget gate determines which information the previous cell state should forget. The input gate controls whether new information should be added to the cell. The output gate determines which information from the updated cell state should be transmitted.

Through these gates, LSTM can learn long-term dependencies. The mechanism of forgetting, retaining, and updating information is learned automatically based on the data during

the training process. This enables LSTM to handle gradient issues more effectively while processing sequential data and model longer-term dependencies.

In the LSTM (Long Short-Term Memory) model, each LSTM cell consists of multiple layers. The basic building blocks of LSTM layers are as follows:

- Cell State
- Forget Gate
- Input Gate
- Output Gate

LSTM networks are a type of RNN architecture specifically designed to model long-term dependencies in sequential data. They have proven to be particularly effective in capturing temporal dynamics and recognizing patterns in time series data. Lip reading by LSTM involves training an LSTM network on visual input, such as video sequences of lip movements, to recognize and interpret spoken words or sentences. By leveraging the sequential nature of lip movements, LSTM networks can effectively capture and utilize contextual information for lip reading tasks.

One of the key advantages of using LSTM networks for lip reading is their ability to handle variable-length input sequences. Unlike traditional methods that rely on fixed-length representations, LSTM networks can process and model temporal information in an adaptive manner, making them well-suited for lip reading tasks where the duration of spoken words or sentences can vary. Moreover, the inherent memory mechanisms in LSTM networks enable them to capture both short-term and long-term dependencies in lip movements, allowing for improved recognition accuracy.

Research in lip reading by LSTM has focused on various aspects, including feature extraction, model architecture, and training strategies. Techniques such as pre-processing lip images, incorporating attention mechanisms, and utilizing multi-modal data have been explored to further enhance the performance of LSTM-based lip reading systems. Furthermore, efforts are being made to create large-scale lip reading datasets that encompass diverse speakers, languages, and environmental conditions to facilitate more comprehensive evaluation and comparison of different approaches.

In conclusion, lip reading by LSTM networks presents a promising approach for improving the accuracy and robustness of lip reading systems. By exploiting the temporal

dynamics of lip movements, LSTM networks can effectively model and recognize spoken language, contributing to advancements in communication accessibility, human-computer interaction, and assistive technologies.

Chen et al. [43], a language-based LSTM model called LipNet is utilized. LipNet is designed for text-level lip reading, aiming to extract lip movements from video inputs and convert them into textual expressions. The end-to-end nature of the model implies that the entire process from input to output is performed within a single model.

Chung et al. [44], demonstrates the successful implementation of lip reading in real-world environments. Using a language-based LSTM model, training is conducted on a large dataset and the model is optimized to accurately predict lip movements. This work highlights the potential of lip reading in real-world applications.

Gan et al [45], focuses on predicting human movements from lip movements using an LSTM-based model. Taking lip movements as input, the model operates as a sequential model to predict human movements. This work showcases the association between lip reading and human dynamics and its potential application in various domains.

Chung et al. [23], sentence-level lip reading accuracy in real-world environments. A language-based LSTM model is employed to develop a system that predicts sentences in different languages from lip movements. This study demonstrates the applicability of lip reading in natural language processing and speech recognition domains.

Ngiam et al. [46] examines multimodal deep learning models. In addition to language-based models, it showcases the utilization of other modalities representing lip movements (audio, visual, etc.). The work emphasizes the importance of combining different data sources to create a more robust and comprehensive lip reading system.

These examples illustrate various aspects and applications of deep learning models in the field of lip reading. Each study focuses on specific objectives such as improving lip reading performance, integrating it with natural language processing, or combining different data modalities.

Chung et al. [47], investigates the potential of lip reading in the field of speaker recognition using deep learning models. Using the VoxCeleb2 dataset, the study demonstrates that lip movements can enhance speaker recognition performance when incorporated into a language-based LSTM model.

Assael et al. [24], performs text-level lip reading using a language-based LSTM model called LipNet. LipNet predicts lip movements from video inputs and converts these predictions into text. The study showcases the applicability of lip reading in natural language processing applications.

Petridis et al. [48], aims to perform visual speech recognition using lip movements. By employing a language-based LSTM model, the study demonstrates that lip movements can enhance the recognition of spoken words. The work showcases the integration potential of lip reading in audio-based speech recognition systems.

Afouras et al. [49], aims to develop a deep learning-based audio-visual speech recognition system by combining audio and lip movements. Using a language-based LSTM model, the study processes audio and visual data together to recognize the words spoken by the speaker. The work demonstrates the effective utilization of lip reading in audio-based speech recognition.

Afouras et al. [50], develops a deep learning-based audio-visual speech recognition model using multiple data streams. By incorporating multiple streams, including audio, lip movements, and linguistic information, the study aims to recognize spoken words. The work highlights the potential of lip reading to improve the performance of audio-based speech recognition.

These examples showcase the various aspects and applications of deep learning models in the field of lip reading. Each study focuses on specific objectives such as enhancing speaker recognition, performing text-level lip reading, improving visual speech recognition, or utilizing multiple data streams for audio-visual speech recognition.

Long Short-Term Memory (LSTM), introduced by Hochreiter and Schmidhuber in 1997, is a type of recurrent neural network (RNN) that has gained significant popularity in the field of sequence modeling and time series analysis. LSTM addresses the limitations of traditional RNNs, such as the vanishing gradient problem, by incorporating a more complex memory mechanism. It is specifically designed to capture long-term dependencies and maintain memory over extended sequences, making it suitable for tasks that require modeling and predicting sequential data.

The key concept behind LSTM is the introduction of memory cells and gating mechanisms, which enable the network to selectively remember or forget information at different time steps. The memory cells act as storage units, retaining information over multiple time steps, while the gating mechanisms regulate the flow of information within the network.

The three main gates in an LSTM unit are the input gate, forget gate, and output gate. The input gate determines which new information should be stored in the memory cells, the forget gate decides which information to discard from the memory cells, and the output gate controls the flow of information from the memory cells to the next layer or output.

The design of LSTM allows it to effectively capture both short-term and long-term dependencies in sequential data. By maintaining a separate memory state and utilizing gating mechanisms, LSTM can learn to selectively update and access information based on its relevance and importance. This makes LSTM particularly suitable for tasks such as speech recognition, language modeling, machine translation, and sentiment analysis, where understanding the context and temporal dependencies is crucial.

In conclusion, LSTM has emerged as a powerful tool in the field of deep learning for sequential data processing. Its ability to capture long-term dependencies and handle vanishing gradients makes it well-suited for a wide range of applications. In the following section, we will present the pseudocode of LSTM, highlighting its key components and operations.

Pseudocode of LSTM;

```
# Load the dataset
train_data, train_labels = load_train_data()
test_data, test_labels = load_test_data()
# Preprocess the data
train_data = preprocess(train_data)
test_data = preprocess(test_data)
# Create the LSTM model
model = create_lstm_model()
# Train the model
model.fit(train_data, train_labels, epochs=10, batch_size=32)
# Evaluate the model
accuracy = model.evaluate(test_data, test_labels)
# Print the results
print("Accuracy: ", accuracy)
```

6. CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional Neural Network (CNN) is a widely used artificial neural network model in the field of deep learning. It provides effective results in visual data analysis, image classification, object detection, and image segmentation tasks, particularly in the field of visual information processing. The success of CNNs stems from deep learning principles that enable automatic learning of data-specific features and specially designed layers.

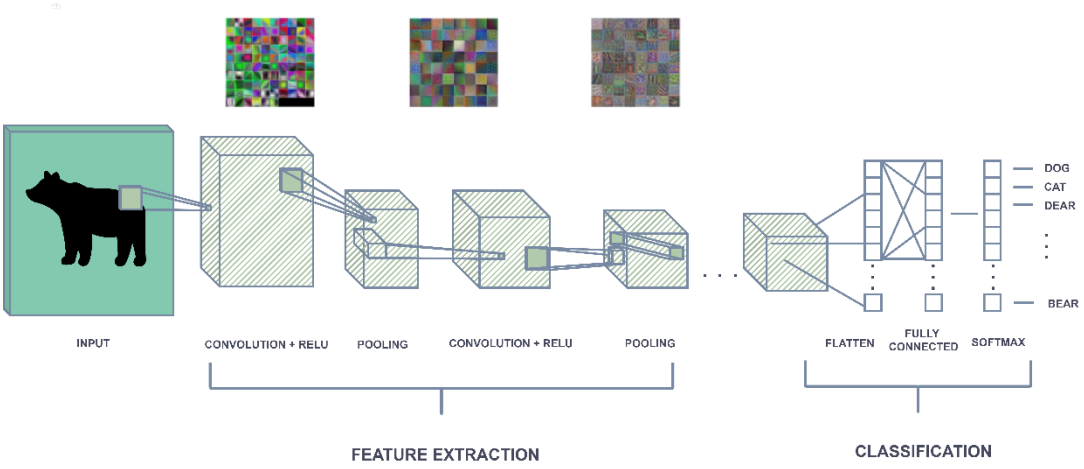

Figure 6.1. Traditional structure of CNN [51]

CNN consists of basic components in figure-5, including an input layer, one or more convolutional layers, activation functions, pooling layers, fully connected layers, and an output layer. CNNs are typically trained on large datasets. The training process involves updating weights to minimize the error (loss) function between input data and target outputs. The backpropagation algorithm is used to calculate the error gradient, and gradient-based optimization methods (e.g., stochastic gradient descent) are applied to update the weights. CNNs have made significant advancements in various fields. In image classification, they have achieved results surpassing human performance in competitions like ImageNet. CNNs are also successfully used in object detection, facial recognition, and research areas. They have potential applications in medicine, automotive, security, robotics, and many other fields.

CNNs have achieved significant success in the field of deep learning and have been particularly effective in visual data analysis. They are known for their ability to automatically learn data-specific features, train on large datasets, and generally perform well. However, factors such as data quality, network architecture, and parameter tuning need to be carefully considered.

Input layer where images or other types of visual data are taken as input. Depending on the size and format of the input data, appropriate resizing and preprocessing steps are performed.

The Input Layer is the first layer of a CNN model and is typically used when processing visual data, such as images. This layer is used to appropriately resize and preprocess the input data based on the size and format of the dataset.

Convolutional layers are the layers where filters are applied to the data to create feature maps. Each filter is used to detect a specific feature on the data. The convolution operation involves sliding the filters over the data with steps determined by parameters like stride and padding.

Convolutional Layers are fundamental components in CNN models and are responsible for creating feature maps by applying filters on visual data. Each filter is used to detect a specific feature in the data. The convolution operation is performed by sliding the filters over the data with steps determined by parameters such as stride and padding.

Activation layers apply non-linear transformations based on the outputs from the convolutional layers. This allows the network to learn more complex relationships and features. Common activation functions used are sigmoid, ReLU (Rectified Linear Unit), and tanh.

Activation Layers are used in CNN models following the Convolutional Layers. These layers process the outputs of the convolutional filters by adding non-linearity and applying activation functions. Activation functions scale the filter outputs and introduce non-linearity, allowing the model to learn more complex relationships.

Pooling layers are used to reduce the size of feature maps and make the features invariant to changes like translation, scale, and rotation. Max pooling and average pooling are commonly used pooling operations.

Fully connected layers are used to transform feature vectors for classification, prediction, or another output format. These layers connect all the features through connections and are typically used in the final layers.

Fully Connected Layers are the final layers in CNN models, responsible for generating outputs such as classification or regression. These layers classify data or make predictions based on the learned features of the CNN.

The output layer produces the final outputs of the CNN model. In classification problems, a softmax activation function is used to obtain a probability distribution, while in regression problems, direct outputs can be produced.

In recent years, the emergence of deep learning, particularly convolutional neural networks (CNNs), has revolutionized the field of lip reading. CNNs have shown remarkable success in various computer vision tasks by automatically learning hierarchical representations from raw input data. This ability to learn and extract discriminative features directly from images makes CNNs a promising approach for lip reading.

The key advantage of CNNs lies in their ability to capture spatial dependencies in visual data. By utilizing convolutional layers, CNNs can effectively extract local patterns and features from lip images, while the pooling layers enable them to capture higher-level representations with spatial invariance. This hierarchical feature extraction enables CNNs to effectively capture the dynamics and variations in lip movements, which are crucial for accurate lip reading.

Moreover, the availability of large-scale lip reading datasets, such as LRW and LRS3, has further fueled the progress in lip reading research. These datasets provide a rich source of labeled lip sequences, allowing researchers to train and evaluate CNN models on large and diverse datasets.

We propose a lip reading system based on convolutional neural networks. We aim to leverage the power of CNNs in extracting spatio-temporal features from lip images and employ deep learning techniques to achieve state-of-the-art performance in lip reading tasks. We will present the architecture of our CNN model, discuss the training process, and evaluate its performance on benchmark datasets.

In conclusion, the application of convolutional neural networks to lip reading has shown great promise in advancing the field. By leveraging the power of deep learning and large-scale datasets, CNNs have the potential to enhance the accuracy and robustness of lip reading systems, paving the way for their practical deployment in real-world scenarios.

The pseudo code provides a general roadmap for deep learning-based lip reading. For a real implementation, the pseudo code may need to be made more specific, and additional deep

learning components (such as pooling layers, dropout, etc.) may need to be included as required. Additionally, the details of the code can vary depending on the programming language used and the deep learning library employed. Therefore, the provided pseudo code serves to provide a general understanding.

Input: Lip image

Output: Recognized word or phoneme

1. Data Preprocessing:
   - Take the lip image
   - Normalize and resize the image
   - Extract relevant image features
2. Define CNN Model:
   - Define the CNN model
   - Create the input layer (based on image size)
   - Specify convolutional layers and activation functions
   - Specify fully connected layers and output layer
3. Training:
   - Prepare training dataset and labels
   - Train the CNN model
   - Update the model using the training dataset
   - Define the loss function and update the network using backpropagation
4. Testing and Prediction:
   - Prepare the test dataset
   - Make predictions using the CNN model
   - Evaluate the predictions (accuracy, precision, etc.)
5. Performance Evaluation:
   - Evaluate the performance of the model
   - Calculate metrics such as accuracy, precision, recall, etc.
   - Analyze the results and gather feedback for improvement steps
6. Prediction with New Images:
   - Use the trained model to make predictions on new lip images
   - Report or utilize the prediction results

The above roadmap illustrates the process of deep learning-based lip reading. Starting from the initial point, the data preprocessing step is performed. Subsequently, a CNN model is defined and trained using the training data. Once the training process is completed, the system proceeds to the testing and prediction phase, where the performance of the model is evaluated. Following the performance evaluation, predictions can be made using new input lip images. The flow chart demonstrates the flow between the lip image input and the recognized word/phoneme output.

This chart provides a simple and comprehensible depiction of the deep learning-based lip reading process. Of course, for a real-world application, a more detailed flow chart could be created, incorporating additional steps or sub-processes to make it more specific.

7. DILATED CNN

Dilated Convolutional Neural Network (Dilated CNN), also known as atrous convolution, is a variant of the traditional convolutional neural network (CNN) architecture that enables increased receptive field without increasing the number of parameters or sacrificing spatial resolution. Dilated CNN has gained significant attention in computer vision tasks, such as image segmentation, object detection, and semantic understanding, due to its ability to capture multi-scale contextual information.

The key idea behind dilated CNN is the introduction of dilation or "hole" in the convolutional filters. Unlike standard convolutional layers, where the filters have a fixed receptive field, dilated convolutions incorporate gaps or "holes" between the filter elements.

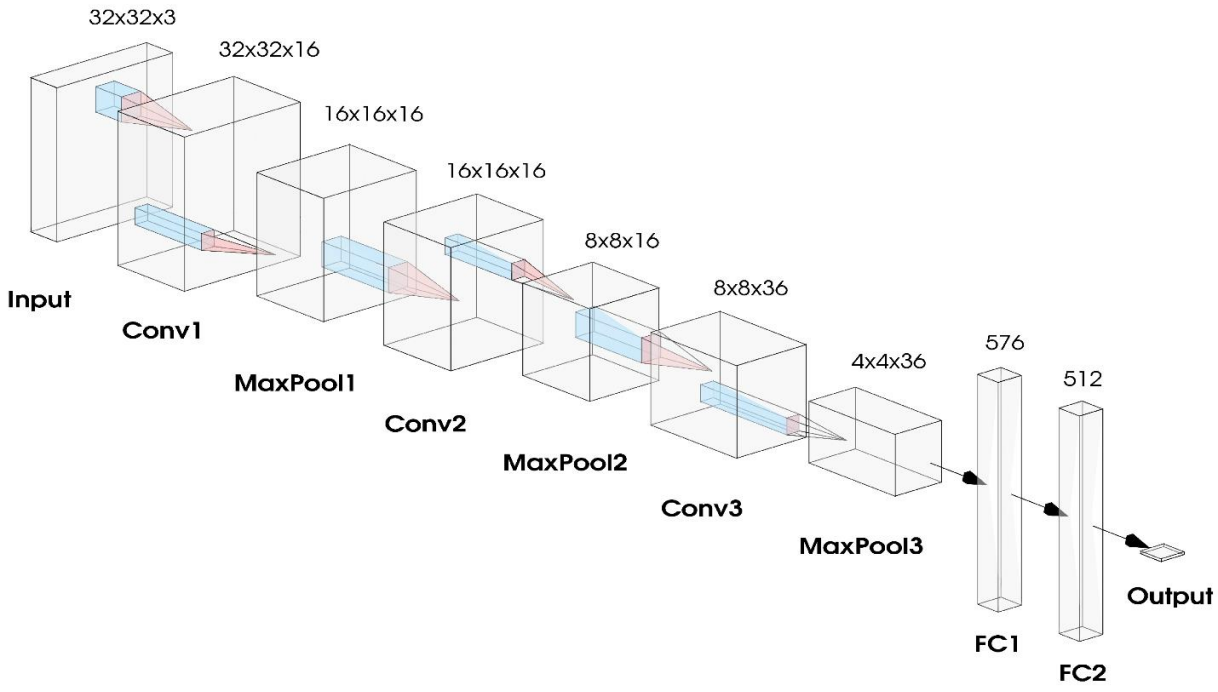

Figure 7.1. Structure of dilated CNN

Dilated CNNs, in figure-6, offer several advantages over traditional CNN architectures. First, they allow for larger receptive fields without increasing the number of parameters, making them more computationally efficient. Second, dilated convolutions preserve the spatial resolution of the feature maps, which is crucial for tasks requiring precise localization or fine-

grained details. Third, dilated CNNs enable multi-scale feature extraction, as they capture information at different levels of granularity due to the varying dilation rates.

The dilation rate determines the spacing between the filter elements and controls the receptive field size. By adjusting the dilation rate, researchers can control the amount of contextual information incorporated into the network. Smaller dilation rates focus on local context, capturing fine details, while larger dilation rates encompass larger context, capturing global structures and relationships.

Dilated CNNs have demonstrated impressive performance in various computer vision tasks. In image segmentation, dilated CNNs can effectively capture both local object boundaries and global contextual information, leading to more accurate and precise segmentation results. In object detection, dilated CNNs enhance the ability to recognize objects of different scales and aspect ratios, improving detection performance. In semantic understanding, dilated CNNs enable better contextual reasoning and modeling, leading to improved understanding and interpretation of complex scenes.

8. THE PRESENTED STUDY

The present study introduces an automated visual speech recognition model based on deep learning, utilizing exclusively visual data. By employing this approach, the computational cost and resource requirements are reduced compared to studies involving multi-modal data. Moreover, this study stands as the pioneering attempt to tackle the lip reading problem within the Ural-Altaic languages, employing a deep learning algorithm on a newly curated dataset. In this thesis in order to recoginize lip reading, we follow steps below;

- Data
    - Data collection and agumentation
    - Data processing
- Optimization
    - CNN based Lip Reading
        - CNN
        - Dilated CNN
    - LSTM based Lip Reading
    - BGRU Based Lip Reading
- Comparision of Concatenated Frame and Discerete Frame in Lip Reading
- Anaysis of results.

8.1. The Improved Dataset

Turkish, when classified based on its linguistic structure, belongs to the family of agglutinative languages. As such, suffixes play a crucial role in determining the meaning of a sentence according to Turkish grammar rules. Additionally, in Turkish, a phenomenon known as liaison occurs when a word starting with a vowel follows a word ending with a consonant. Liaison refers to the effect produced when these two letters are connected and read together, and it significantly impacts the meaning of the sentence. For example, the phrases "top aldı" (bought a ball) and "topaldı" (was lame) have distinct meanings due to liaisons in the letters "p" and "a," despite having the same letter order.

In the existing literature, several datasets have been created using various methods for lip reading studies. However, it has been observed that no specific lip reading dataset for the

Turkish language has been developed, except for the dataset presented by Atila and Sabaz [53]. These authors created two new datasets using image processing techniques, one consisting of words and the other consisting of sentences. The sentence dataset includes classes such as "Which department did you get?" and "May I help you?", while the word dataset includes classes like "Programmer" and "Video". One notable distinction between their dataset and ours is that all the words and sentences were created under the same environmental and lighting conditions. In contrast, our dataset was obtained from a diverse range of YouTube videos, resulting in hundreds of different speaking profiles.

In another relevant study conducted by Matthews et al. [54], they developed a customized audio-visual (AV) database called AVLetter, which consisted of isolated letters. The dataset encompassed three repetitions of all the letters in the alphabet, spoken by ten different speakers (including five males, two of whom had mustaches, and five females), resulting in a total of 780 utterances. The researchers employed various techniques, including internal and external contour methods, along with a novel bottom-up approach that involved extracting features directly from pixel intensity using nonlinear scale space analysis. Furthermore, they trained a Hidden Markov Model (HMM) and obtained an accuracy score of 44.6%.

In another work [55], two new datasets were introduced and publicly released: LRS2-BBC [56], which includes thousands of natural phrases from British television, and LRS3-TED [57], containing hundreds of excerpts from over 400 hours of TED and TEDx videos [58]. These datasets encompass unrestricted natural language sentences and videos featuring different individuals, unlike synthetic datasets generated with controlled background, lighting, and angle conditions. Researchers have demonstrated that combining visual speech recognition (VSR) and auditory speech recognition methods, particularly in the presence of vocal noise, leads to significant improvements in lip reading studies.

Yang et al. [59] introduced a large-scale benchmark dataset called LRW-1000, specifically designed for lip reading research. This dataset consisted of 1000 classes, encompassing 718,018 samples from over 2000 speakers. Each class represented syllables of Mandarin words, composed of one or more Chinese characters. The LRW-1000 dataset was carefully curated to emulate real-world conditions, exhibiting significant variations in various aspects, such as the number of samples per class, video resolution, lighting conditions, and speaker characteristics, including pose, age, gender, and makeup.

In a similar vein, Egorov et al. [60] constructed a Russian lip reading dataset known as LRWR. This dataset consisted of 235 classes and involved 135 speakers. The authors provided a detailed description of their dataset aggregation pipeline and presented comprehensive statistics in their paper. By creating a large-scale Russian dataset, they contributed to the visual lip reading dataset research, which has been predominantly focused on English language lip reading studies.

Chung and Zisserman [61] pursued the goal of word recognition solely based on visual cues from a speaking face, without utilizing phonetic information. They developed an automated data collection pipeline from TV broadcasts, resulting in a dataset containing over a million examples of spoken words by different individuals.

In summary, these studies highlight the creation of diverse and sizable lip reading datasets, such as LRW-1000, LRWR, and the dataset generated by Chung and Zisserman. These datasets facilitate research in lip reading by incorporating real-world conditions, encompassing a wide range of linguistic and visual variations, and expanding beyond the dominance of English language studies in the field.

A two-stream convolutional neural network was developed to learn the correlation between audio and visual mouth movements from unlabeled data. The training results achieved with this dataset and model surpassed the performance of publicly available datasets, namely Columbia [62] and OuluVS2 [63].

Anina et al. [64] presented the OuluVS2 dataset, which was specifically aggregated for analyzing non-rigid mouth movements. This dataset comprises recordings of more than 50 speakers uttering English phrases, numbers, three-word phrases, and three sentences. The dataset includes thousands of videos captured simultaneously from five different viewing angles, ranging from frontal to profile views. An HMM-based visual speech recognition (VSR) system was developed and tested on the OuluVS2 dataset. The recognition results revealed that the 60° angle provided the highest accuracy score of 46%, whereas the score was 42% for the 90° angle (front view).

The Arabic Visual Speech Dataset (AVSD) [65] consists of 1100 videos containing recordings of 10 daily communication words, such as hello, welcome, and sorry. The dataset was collected from 22 speakers under realistic conditions, including various indoor rooms with

different lighting conditions. VSR experiments were performed on the AVSD using a support vector machine (SVM), and the algorithm achieved an average word recognition rate of 70.09%.

Sujatha and Krishnan [66] compiled a dataset involving 10 participants who recorded stable ambient conditions while uttering 35 different words. For training, 4900 samples were collected, with each of the 7 participants pronouncing 20 samples for each word. Additionally, 2100 samples were used for testing, with each of the 3 participants providing 20 samples for each word. The videos of the participants were processed using a face localization module to detect the facial region, and subsequently, the mouth region was determined.

In summary, these studies demonstrate the creation and utilization of various datasets for analyzing visual speech and mouth movements. These datasets include OuluVS2, AVSD, and the dataset prepared by Sujatha and Krishnan. The experiments conducted on these datasets, employing different recognition algorithms and evaluation metrics, contribute to advancing the field of visual speech analysis and recognition.

In reference [67], a dataset was created and utilized to address lip reading challenges, incorporating audio and lip movement data from various videos containing readings of identical words such as "book," "come," and "read." The proposed method employed the VGG16 pre-trained convolutional neural network (CNN) architecture for data classification and recognition. The recommended model achieved an accuracy of 76% in visual speech recognition (VSR).

In their work, Xu et al. [68] utilized multi-expansion temporal convolutional networks (MD-TCN) for the purpose of word prediction in lip reading tasks. Their methodology involved incorporating a self-attention block following each convolutional layer to augment the model's classification and scanning capabilities. By evaluating their approach on the LRW dataset (69), they achieved an accuracy of 85%, thus showcasing a marginal improvement of 0.2% compared to other networks with similar architectures [70].

Berkol et al. [71] conducted a comparative analysis using the dataset introduced in this study to assess the performance of the dilated convolutional neural network (DCNN) model against the convolutional neural network (CNN) model utilized in their prior research. The multiclass classification model yielded a test accuracy of 59.80% for the DCNN, whereas the CNN model in their earlier study achieved an accuracy of 72%. It was observed that the CNN outperformed the DCNN in terms of both time and accuracy. The relatively lower accuracy score of the DCNN model can be attributed to the utilization of a non-synthetic dataset with

intricate features, posing challenges for the model. Existing lip reading datasets employed in prior studies [72, 73] were primarily obtained under controlled conditions. The contribution of our study to the existing literature lies in the provision of a non-synthetic Turkish lip reading dataset, which, to the best of our knowledge, represents the first of its kind. This dataset was derived from natural speech recordings, with careful examination of the videos to eliminate any factors that might impede accurate lip movement analysis, such as the presence of microphones, subtitles, or occluding hands. The data exclusively focused on capturing facial expressions for the purpose of lip movement analysis.

Overall, these studies highlight the development of lip reading datasets, utilization of pre-trained CNN architectures, and the exploration of novel approaches such as MD-TCN. Additionally, the dataset proposed in this study contributes to the advancement of Turkish lip reading research by providing a non-synthetic dataset obtained from natural speech recordings. However, this dataset, consisting of wide-framed images capturing people pronouncing various words, can be utilized for different research problems with appropriate data arrangements. It facilitates the development of word or phrase recognition from a speaking face without audio [74], without relying on lip-motion recognition.

8.1.1. Dataset collection

The data collection process commenced by identifying relevant YouTube videos containing the specified words. Screen recording techniques were employed to capture the videos. Throughout the data collection phase, particular emphasis was placed on creating a diverse sample set, encompassing variations in gender (male/female), age groups (adult/child/elderly), indoor/outdoor settings, lighting conditions (light/dark), presence/absence of mustache, presence/absence of makeup, and slight variations in face angles.

Due to these data collected for the lip reading problem are obtained from the videos of the speakers who continue in their natural flow, the images are challenging in terms of diversity (see Fig. 39). In some cases, speakers do not turn their face directly to the camera. Furthermore, there are situations such as light differences in the image, image quality, and the speaker being far away. In addition to these, there is also a problem that creates personal diversity such as objects such as microphones coming in front of the speaker in the images obtained, the speaker's mustache and lipstick.

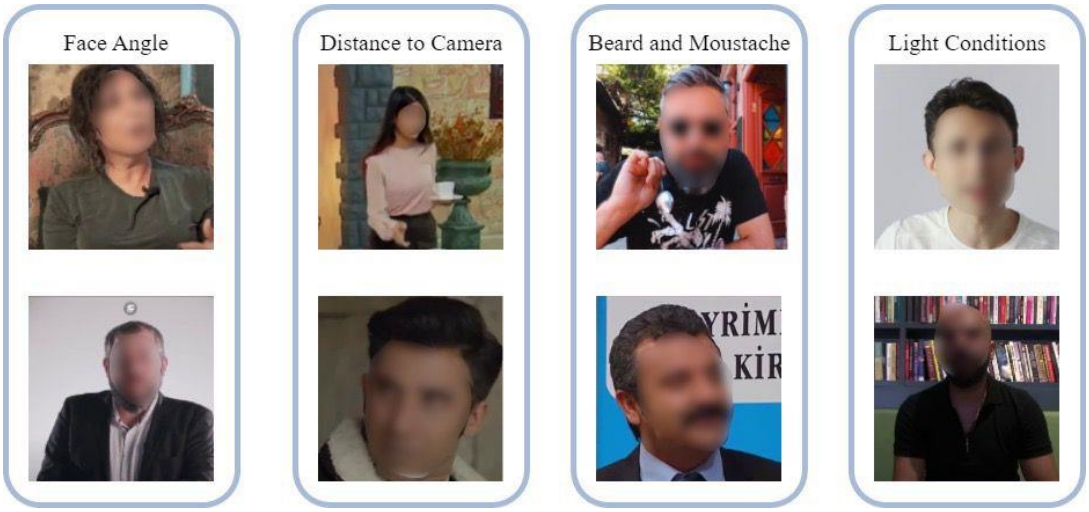

Figure 8.1. Data challenges

8.1.2. Frame extraction from videos

After collecting 2335 instances, they were segmented into frames using the Python library, OpenCV. During the frame extraction process, a script was developed to identify the specific second at which each word started and determine the video's frames per second (fps). Subsequently, frames captured within a 2-second interval following the identified second were extracted and saved as individual images. The resulting images varied based on the fps value. Generally, since the videos were recorded at a standard fps rate of 30, a total of 60 frames were obtained for each 2-second block.

Firstly, it was crucial to create a balanced multi-class dataset. Working with a balanced dataset in terms of labels reduces challenges and allows developers and researchers to focus on developing more optimal and diverse models. In this study, we placed great emphasis on obtaining an approximately equal amount of data for each label. Table 1 provides the number of samples available for each class in the dataset.

Secondly, ensuring a normal distribution of frame numbers for these words in the dataset is crucial for training high-performance machine learning models. Since the difference in the number of examples for each class instance is minimal, the model's results will exhibit consistent performance.

Table 8.1. Number of instances in the dataset.

Words and Phrases	Number of Instances
başla (start)	225
bitir (finish)	244
merhaba (hello)	268
günaydın (good morning)	232
selam (hi)	235
hoş geldiniz (welcome)	226
özür dilerim (sorry)	209
görüşmek üzere (see you)	224
afiyet olsun (enjoy your meal)	235
teşekkür ederim (thank you)	237

Secondly, apart from the relative frequency of each class, the number of frames associated with each word is a critical aspect in machine learning models, particularly in deep learning. It can serve as an influential parameter in real-time word recognition. Figure 7 illustrates the distribution of each class based on the number of frames. The top five labels correspond to phrases such as "teşekkür ederim" and "hoş geldiniz," while the remaining labels represent individual words like "günaydın" and "selam." The number of frames for words ranged approximately between 3 and 26, while for phrases, it ranged between approximately 7 and 33.

To analyze the distributions of frame numbers, the Pandas skew() method, which provides unbiased skew values, was employed. The skewness coefficients for the words "günaydın," "merhaba," "selam," "başla," and "bitir" were 0.06, 1.46, 0.86, 0.09, and 0.54, respectively. For the phrases "afiyet olsun," "görüşmek üzere," "hoş geldiniz," "özür dilerim," and "teşekkür ederim," the skewness coefficients were 0.10, -0.16, 0.07, 0.48, and 0.72, respectively. High skewness coefficients were observed for the words "selam," "merhaba," and "teşekkür ederim," indicating right-skewed distributions, while the skewness coefficients for other words and phrases were close to 0, indicating normal distributions.

For the word "merhaba," the mean frame number was 12.7, the median was 12, and the mode was 10, indicating a normal distribution. Similarly, for "günaydın," the mean, mode, and median values were 9.1, 9, and 9, respectively, indicating a normal distribution. The presence

of children's songs among the videos used for the word "merhaba" resulted in slower speech compared to other recordings. The non-normal distributions observed for certain classes indicated a greater diversity among speakers and the inclusion of various video types, such as vlogs, TV series, or clips, in our dataset.

Understanding and accessing the dataset is facilitated by familiarity with its directory structure. The directory hierarchy is organized as follows: the top-level directory corresponds to specific word or phrase tags, such as "başla" or "teşekkür ederim". Within each word folder, there are subdirectories representing individual instances, which are sequentially named using three-digit numbering. The final level of the dataset architecture comprises processed frames extracted from the corresponding videos, and these frames are sequentially named using two-digit numbering, such as "01.jpg, 02.jpg, ..., 28.jpg". Figure 9 illustrates the hierarchical structure of the dataset directories.

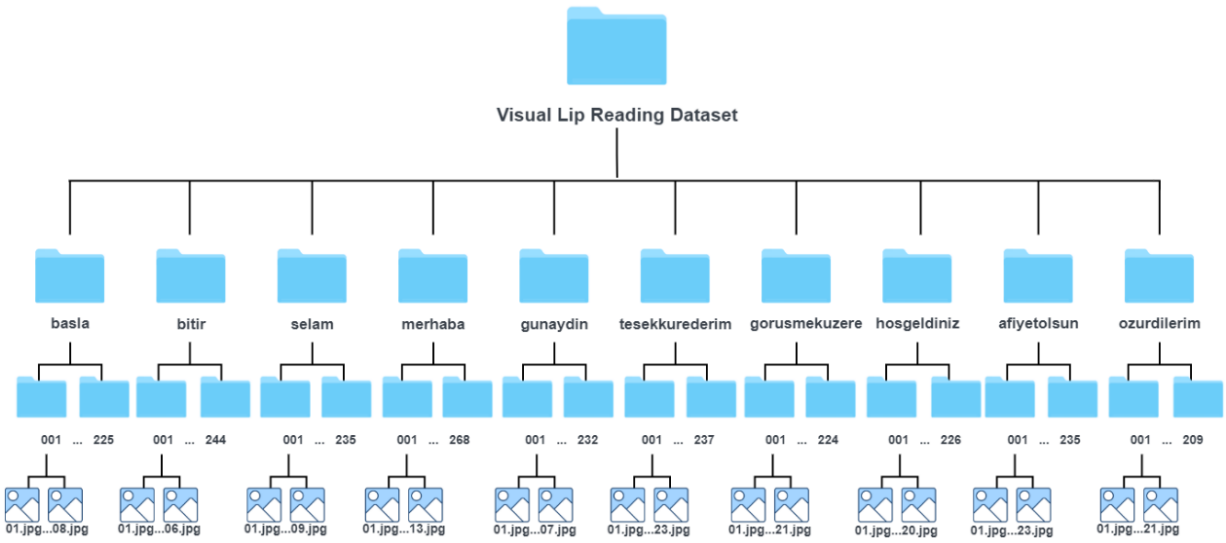

Figure 8.2. The directory architecture of the dataset; “merhaba” (hello), “selam” (hi), “başla” (start), “bitir” (finish), “günaydın” (good morning), “teşekkür ederim” (thank you), “hoş geldiniz” (welcome), “görüşmek üzere” (see you), “özür dilerim” (sorry), and “afiyet olsun” (enjoy your meal) are words and phrases that appeared in the first step. A subdirectory has samples of words and phrases contained within it. The last step of the architecture shows the frames of the related word.

8.1.3. Frame cropping

To facilitate the identification of the person speaking and enable accurate lip reading, frames containing multiple human faces were deemed complex and challenging. As a result, a manual cropping process was conducted using image cropping applications to exclude images with multiple faces, except for the face of the person of interest. Care was taken to ensure that the entire face of the speaker, with clear visibility of lip movements, remained within the field of view during the cropping process. Frames with no other faces, obstructions, or profile views that hindered lip movement were selected for inclusion in the dataset. The emphasis was on preserving the background and obtaining real-world instances without removing inherent noise during the cropping process.

A review of previous studies revealed only one dataset related to Turkish lip reading. What sets this study apart from that dataset, where all words and sentences were created under controlled ambient and lighting conditions, is that it introduces a non-synthetic lip reading dataset that had not been previously developed. The data collection methods employed in the two studies differed significantly. While the previous dataset was generated by 24 speakers who specifically pronounced certain words and phrases, our dataset captured the moments in which the relevant word was spoken from various people's YouTube videos. Additionally, the pronunciation of words in the Turkish language is influenced by various factors, such as the speaker's accent, the presence of liaison, and word stress. Thus, the aim was to create a dataset suitable for real-life conditions by collecting samples from a diverse range of individuals.

The dataset we created contributes to visual lip reading studies and enables researchers to produce more realistic results due to the complex environmental conditions encountered in real-life scenarios. By utilizing this dataset in lip reading studies, researchers can contribute to solving forensic cases, enhancing the lives of hearing-impaired individuals, and introducing innovative approaches to language education. The dataset focuses solely on capturing facial expressions to describe lip movements. However, due to its wide-framed images of individuals pronouncing various words, it holds the potential to be utilized in addressing various research problems after appropriate data adjustments.

In summary, the manual cropping process was conducted to exclude frames with multiple faces, and the dataset created in this study stands out as a non-synthetic lip reading dataset that

captures real-life conditions. Its potential applications extend beyond lip reading studies, making it a valuable resource for diverse research endeavors.

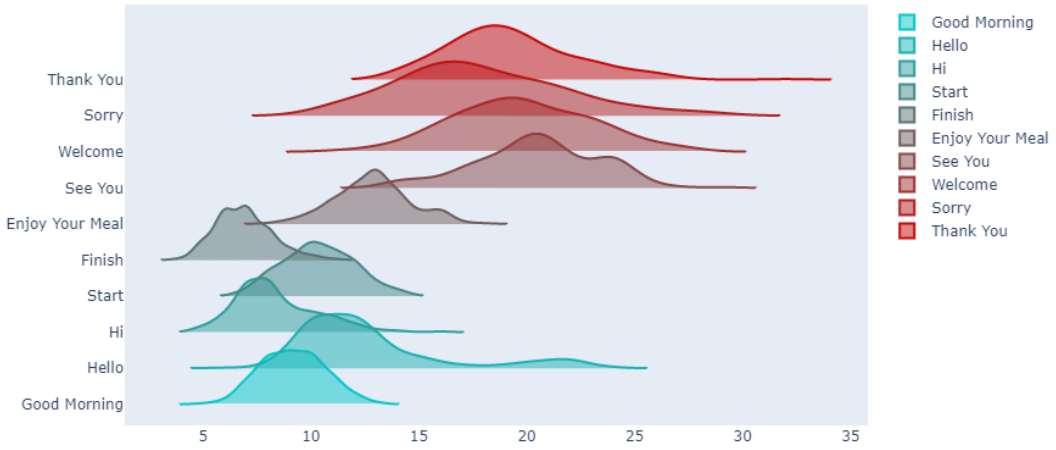

Figure 8.3. Frame number distribution for each word such as “hello” (merhaba), “hi” (selam), “start” (başla), “finish” (bitir), and “good morning” (günaydın) and phrases such as “thank you” (teşekkür ederim), “welcome” (hoş geldiniz), “see you” (görüşmek üzere), “sorry” (özür dilerim), and “enjoy your meal” (afiyet olsun).

Lastly, a correlation matrix was generated to explore potential linear relationships between the classes. The following steps were followed to identify causal or non-causal relationships:

Firstly, clear and representative examples were selected from the dataset for each class, ensuring accuracy in the results. The lips were then extracted from the original images since analyzing lip movements is crucial and enables working with reduced data.

Next, the sequence of arrays was flattened to a one-dimensional summarized array by computing the median value for each index of the images. The finalized arrays for each class were subjected to the Pearson correlation method. The Pearson correlation coefficient ranges from -1 to 1. A value close to 1 indicates a positive relationship between the variables, suggesting a positive causal relationship. Conversely, a value close to -1 indicates a negative causal relationship between the variables. If the value is closer to 0, from both the negative and positive sides, it suggests a non-causal relationship between the variables, indicating no linear relationship.

In summary, by applying the Pearson correlation method to the flattened arrays, we examined the presence of linear relationships between the classes. The correlation matrix

provides insights into the nature of the relationships, helping to identify potential causal or non-causal associations among the variables. In Figure 2, we depicted the Pearson correlation using a heatmap, which revealed the correlation patterns among different classes. It is evident that certain classes exhibited high positive correlations. For instance, "afiyet olsun" and "günaydın" displayed a strong positive correlation, with a correlation coefficient of approximately 0.9. Similarly, the classes "merhaba" and "başla" demonstrated a positive correlation, albeit of lesser strength, with a correlation coefficient of around 0.6. Nevertheless, we did not observe a substantial overall relationship between the classes. Furthermore, no significant negative correlations were observed, unlike the strong positive examples mentioned earlier. It is worth noting that the dataset proves to be valuable for addressing classification problems since the patterns exhibited by different classes are distinct and amenable to various methods, including deep learning and machine learning algorithms.

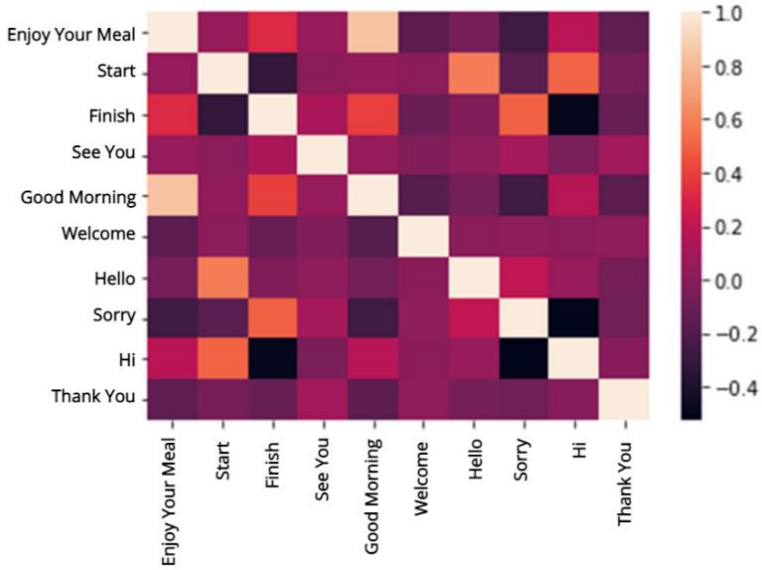

Figure 8.4. Distance matrix for each class such as “hello” (merhaba), “hi” (selam), “start” (başla), “finish” (bitir), “good morning” (günaydın), “thank you” (teşekkür ederim), “welcome” (hoş geldiniz), “see you” (görüşmek üzere), ”sorry” (özür dilerim), and “enjoy your meal” (afiyet olsun) based on the image features.

Creating a distance matrix for a word dataset is done to evaluate word similarities and relationships. The distance matrix contains the measure of similarity or distance between each

word and all other words. Distance metrics are commonly used to obtain a quantitative assessment of semantic or lexical similarities between word pairs.

There are various methods to create a distance matrix. One approach is to use word vector representations and compute similarity metrics. For example, word embedding models can be used to generate word vectors, and then distances between these vectors can be calculated to create the distance matrix. This matrix can be utilized to measure word similarities or relationships.

A distance matrix can be useful in various natural language processing (NLP) tasks such as word classification, word clustering, and analyzing word relationships. It can be used to discover similar words or meaningful word groups, explore word relationships, or perform semantic searches at the word level.

For these reasons, creating a distance matrix for a word dataset is a common approach to analyze relationships between words and obtain similarity measures.

8.1.4. Detection of Lip

In the lip-reading problem, the RGB images are not important for the continuity of the studies. Images are converted to gray scale in order to reduce computational and time costs in face and lip detection studies and later during deep learning model training.

First, we cut the faces from the human images we collected using the dlib library, which is a ready-made library, since the faces on the images need to be handled. The get_frontal_face_detector() function we use does not receive face detection without taking any parameters. When this function is called, it returns the pre-trained HOG+Linear SVM face detector of the dlib. HOG+LINEAR SVM works fast and effectively. Due to the nature of the HOG, it adapts to rotation and viewing angle situations. This detector is built using a histogram of oriented gradients (HOG) and a linear SVM. It is suitable method for real-time face detection due to its its rapid detection.

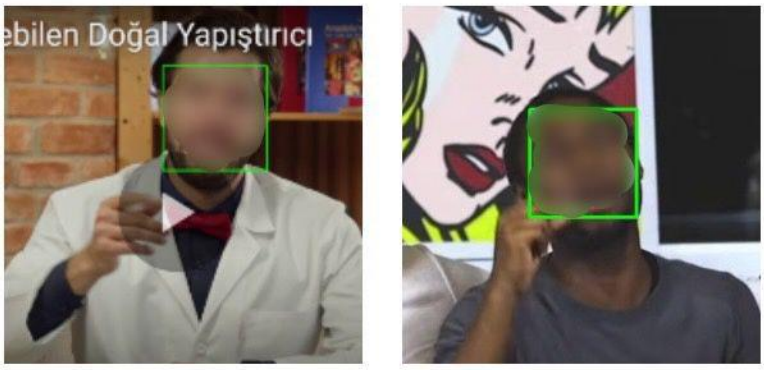

Figure 8.5. Face detection with HOG+SVM

As can be seen in Fig. 40, even if the faces are angled or if there is an obstacle in front of the face, an accurate face detection can be made, including the lips.

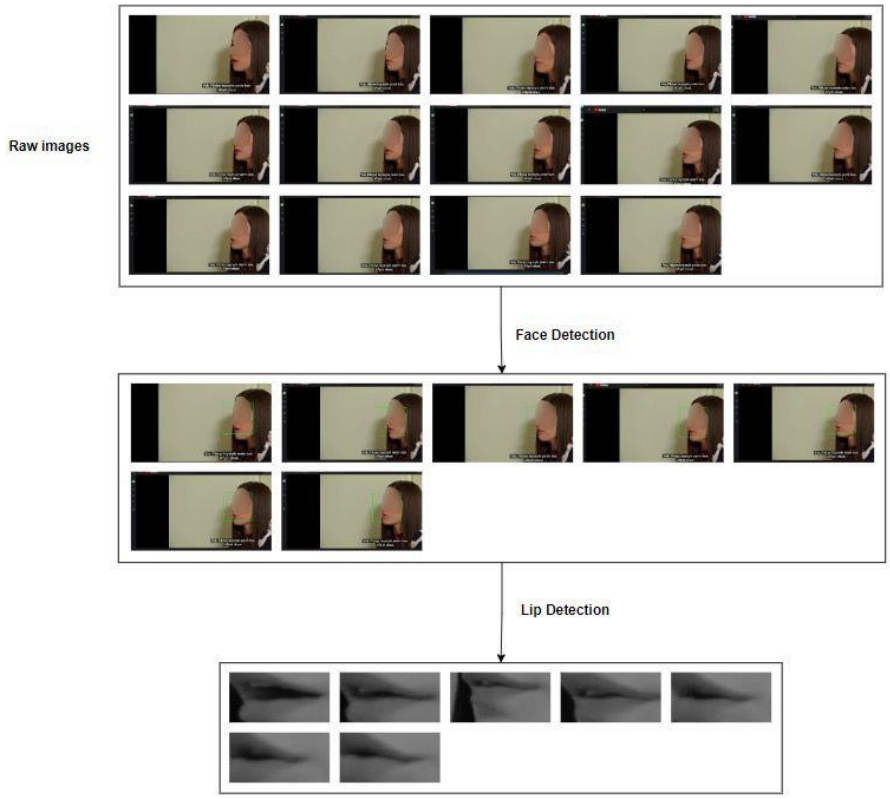

Figure 8.6. Lip Detection

In lip-cutting studies, using the OpenCV library, the contour of the relevant region is drawn by specifying a series of points to take the lip part. Since the 49-68 range corresponds to

the lip region in the landmarks, the relevant range on the obtained face image is cut. Then, with the help of the boundingRect() function, a rectangular image of the determined region is taken. Fig. 41 shows firstly, the original raw images, the faces detected in the second step, and the cut lip images at the end. Lip detection is also less than the number of raw images, as there is no corresponding face detection for each raw image. Although there are similar images in terms of angle and light in each image, it was observed that face detection could not be performed for each of them.

Finally, the lip images obtained are recorded in 100X200 size to be used in the next steps.

Total dataset information seen in table 8.2. below;

Table 8.2. Total Dataset information

Specification	Details
Data Source	YouGlish website
Total Number of YouTube Videos	7000+
Language	Not specified
Word Count	100 words
Phrase Count	100 phrases
Frame Count	20000+ frames
Gender Distribution	Approximately 50% male, 50% female
Age Distribution	Adult, child, old
Environment Distribution	Approximately 25% outdoor, 75% indoor
Lighting Distribution	Approximately 50% light, 50% dark
Facial Features	Variation in mustache presence, makeup presence
Face Position	Slight angle

8.1.5. Lip representation

In the first approach developed, each sequential image of a sample is used in the deep learning model so that the flow is preserved. As a second approach, 15 images are combined and used as a single smaller image. After the concatenation process, each 100x200 image is resized to 20x40 in order not to obtain a very large image. If the frame number of the relevant sample of the lip is less than 15, an image filled with 0 values on the gray scale is added. If it is more than 15 it is removed. When 15 frames are sequentially combined as 3 rows and 5 columns,

a 60x200 image is obtained. In Fig. 8.7. shows 15 sequence images produced in combination. In the case of separate lips, these images are used as a series of 15 images, providing a stream instead of a single image.

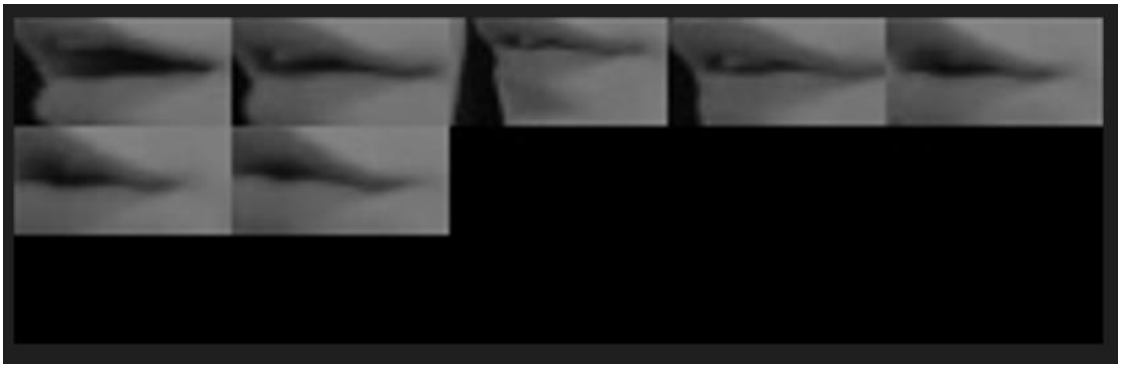

Figure 8.7. Concatenated frame mouths

To ensure the accuracy of the dataset, frames were selected based on the moment when the specified word was first spoken, minimizing the inclusion of lip movements from other words within the same video. Subsequently, using a simple code, the frames were converted into images for extraction, as detailed in the subsequent section.

During the screen recording process, certain videos were eliminated if they did not adequately capture the lip image or if other objects obstructed the view. Examples of such situations included hand movements obstructing the face, instances where the lip image temporarily moved out of the field of view, or default subtitles covering the lip movements.

8.1.6. Data augmentation

Data augmentation techniques are used when the dataset size is not enough to train deep learning algorithms or when the data quality or variety is not enough. With the help of augmentation techniques, classification results can be enhanced. In this work, we applied three different augmentation techniques to the dataset. It is important to note that augmentation techniques were implemented for the whole dataset since the visual lip reading problem concerns the sequence data where data are all images. The first augmentation technique is a horizontal flip (see Figure 14, the second row). A horizontal flip is a mirror reflection by the y-axis. The second augmentation technique is inverting the image by subtracting pixel values from

255 (see Figure 14 the third row). The last augmentation technique is sigmoid contrast (see Figure 8.8, the last row). The technique is applied with the sigmoid function in Equation (1), where the gain is (5, 10) and the cutoff is (0.4, 0.6). After applying the augmentation techniques, the dataset size expanded from 1390 to 5560 sets of examples.

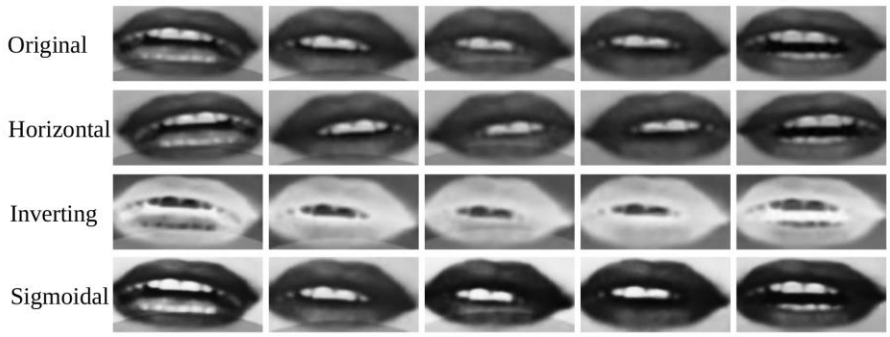

Figure 8.8. Data augmentation techniques applied on visual lip reading in Turkish dataset.

$$f(v) = 255 \times \frac{1}{(1 + exp^{(gain \times cutoff - \frac{v}{255})})} \tag{8.1}$$

8.2. Our Study with Deep Learning Models

Visual speech information is critical when voice data is noisy, difficult to acquire, or lacking context. People find it extremely difficult to understand what someone is saying merely by watching their mouth motions [75]. For instance, adults who are deaf or hard of hearing only achieve an accuracy of approximately 17% for a limited sample of 30 monosyllabic words and approximately 21% for 30 complicated [76].

In addition to understanding or recognizing the words and phrases by the listener as a research question, lip reading can be applicable to many areas in the industry, such as information security [77, 78] speech recognition [79, 80, 81], and driver assistance [82]. Moreover, it gives people with hearing problems a new way to communicate with the outside world [83, 84]. Regular people who do not have hearing problems can also benefit from lip reading in settings where speaking aloud is improper, such as a meeting room [85]. Lip reading has recently been used as a novel biometric identification method for mobile devices [86]. As a result, lip reading and its applications are inseparable from society.

Lip reading models which use multi-model data are widely used in the research field, e.g., Chung et al. [87] and Iwano et al.[88]. Despite the advantages of working with multi-model data, there are significant drawbacks, such as separating noise from data captured from crowded environments and requiring higher data storage, which also affects the model training efforts. Furthermore, even while voice-image-based lip reading has shown its usefulness, only-image-based lip reading demonstrated good results as well Fenghour et al.[89]; Pandey and Arif [90]. However, a challenging problem, distinguishing similar lip movements for different words or phrases, reveals itself when the dataset contains only-image data. Distinguishing sounds with similar lip movements is a challenging problem. Additionally, since the algorithm can handle one person's data, it can be challenging to decide who is talking and whom the algorithm will take into account when there is more than one person on the camera screen. However, it is still easier to preprocess image data.

The Turkish lip reading model is trained and tested on only-image based dataset to increase the classification success rate for various deep learning models, which are Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), and Bidirectional Gated Recurrent Units (BGRU). The following sections cover the data preprocessing stages and the modeling experiments in detail.

Artificial Intelligence (AI) researchers have recently become interested in the lip reading problem. Each language has a different structure since lip reading is sensitive in terms of language and sound. Because of that reason, there are various works for some languages [91, 92]. Additionally, there are a ton of state-of-the-art studies available in terms of data types and languages. Some important models and approaches are as follows.

Conventional approaches typically rely on handcrafted features, which are too complicated and time-consuming to train neural networks. The images are converted into numerical features that can be fed into deep learning algorithms for classification. Haq et al. [92] used both visual and sound data to train the model, a combination of a spatiotemporal convolution layer and SE-ResNet-18 network with a BGRU back-end, 1D convolutional layer and fully connected layers performed on Daily Mandarin Conversation Lip Reading dataset.

The tiny and intricate signal patterns created by mouth motion are well captured by the data collection approach developed by Zhang et al. [93]. The authors also suggest a set of algorithms to extract signal profiles linked to mouth motions and reduce interference factors

like multi-path. Then, to improve the recognition accuracy at the word level, a carefully crafted set of features, including time-domain statistical and frequency-domain features, are retrieved from the signal. A transfer-learning-based strategy is utilized to improve the model's robustness in cross-environment situations and lower training costs when employed in a new environment. Peng et al. [94] suggest a network with channel-temporal attention and deformable 3D convolution, where channel-temporal attention takes advantage of the inherent correlation of features to force the network to focus on necessary keyframes, and deformable 3D convolution adapts the sample position adaptively based on the lip architecture.

Xue et al. [95] propose a complete Bayesian learning approach to account for the underlying uncertainty in LSTM-RNN and Transformer Language Models (LMs). LSTM-RNN or Transformer LMs are used to model the uncertainty surrounding their model parameters, choice of neural activations, and hidden output representations. In order to automatically choose the best network internal components for Bayesian learning utilizing neural architecture search, efficient inference methods were applied. Additionally, a minimum of one sample of a Monte Carlo parameter was used. These make it possible to reduce the computing expenses associated with Bayesian NNLM training and evaluation.

Fenghour et al. [96] wrote a valuable survey for contrasting different approaches concentrating on neural networks and feature extraction. The authors' key finding is that Attention-Transformers and Temporal Convolutional Networks benefit from Recurrent Neural Networks. They concentrate on both audiovisual and merely visual information. Additionally, they mentioned letter-based, word-based, and sentence-based approaches that applied to English, Chinese, German, and Arabic, among other languages. From a different perspective, data augmentation techniques such as "salt and paper", "gaussian", and "speckle" noise adding, and "median" filtering were used to increase the dataset size (Ozcan and Basturk [97]). Moreover, they used AlexNet and GoogleNet pre-trained CNNs on the AvLetters dataset.

For improved accuracy, the Haar Feature-Based Cascade classifier and CNN network are utilized [98]. According to [99], there exist several studies focused on enhancing accuracy in the field. In these works, the authors emphasized the importance of geometric details, such as mouth height, width, and area. For the purpose of recognition, a Hidden Markov Model (HMM) was utilized as a challenge. Another application that used articulated feature extraction approaches used a dynamic Bayesian network for recognizing short phases, and a support vector

machine for classification [100]. HMM is another application that leverages geometric information from the side-face. Lip-contour geometric features are the angles formed by two lines taken from upper and lower lip locations (LCGFs). As LCGF steps, the authors identify a lip area, extract a lip center point, and determine lip lines and a lip angle. [101] is a favorable survey for comparing different approaches, especially neural networks and feature extraction. The authors' main conclusion is that Attention-Transformers and Temporal Convolutional Networks have benefits over Recurrent Neural Networks. They concentrate on both audiovisual and only-visual information. They also included letterbased, word-based, and sentence-based approaches that apply to English, Arabic, Chinese, and German. In [102], the authors utilized pre-trained Convolutional Neural Networks (CNNs) such as AlexNet and GoogleNet on the AvLetters dataset. To expand the dataset, data augmentation techniques were employed. These techniques involved adding noise through "gaussian," "salt and pepper," and "speckle" filtering, as well as applying sharpening using "unsharp" and softening using "median" filtering.

In [103], a CNN was introduced as a novel network for digit classification. The dataset consisted of numbers ranging from 0 to 9, spoken by three female and three male speakers and repeated up to 100 times. The VGG19 network was employed to capture spatial characteristics, while the Attention-based Long Short-Term Memory (LSTM) network was used to capture temporal characteristics. An alternative approach to LSTM is the use of Temporal Convolutional Networks [104In [105], the authors propose a Multi-Scale Temporal Convolution approach for word-level classification. They conducted experiments using data consisting of only audio, audio-visual, and only visual modalities. In [106], a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks was employed for classification. The authors utilized a VGGNet pre-trained on human faces of celebrities from IMDB and Google Images. They contributed by concatenating images and extracting temporal information using LSTM. To facilitate the learning of mapping mouth movements to characters, [107] introduces the "Watch, Listen, Attend, and Spell" (WLAS) network. This network aims to duplicate videos of mouth movements and convert them into corresponding characters or words.WLAS includes WAS, which is a model that only works with photos. They also proposed a curriculum-based learning technique to cut down on training time and reduce overfitting. Additionally, for visual speech recognition applications, the "Lip Reading Sentences" (LRS) dataset was published, which comprises over 100,000 natural sentences from British television.

LipNet was designed and trained for end-to-end sentence and phrase-level predictions. The proposed model in [108] utilizes spatiotemporal Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and the connectionist temporal classification (CTC) loss for character-level prediction. The authors conducted their research using the GRID corpus dataset, which is a publicly available dataset annotated at the sentence level [109]. Another model, called LipType, was developed to achieve advanced speed and accuracy. The authors also focused on improving the model's performance in low-light conditions. The model consists of multiple stages. In the first stage, a spatiotemporal feature extraction method is employed, which includes facial landmark correction using Kalman Filtering, 3D-CNN, and 2D SE-ResNet. The outputs from this stage are then fed into Bidirectional Gated Recurrent Neural Networks (RNNs) with the CTC loss function for further processing and prediction.Fernandez-Lopez and Sukno et al. [98] stated that they use digits or letters and words or sentences as targets for the problem. They developed an end-to-end algorithm dominated by RNNs, and achieved approximately 40% advancement in the word prediction rates. The algorithm, developed by Fenghour et al. [96], only uses visual signals and lacks language. Visemes in continuous speech is recognized using a uniquely developed transformer with a unique topology. The use of perplexity analysis to translate visemes into words. Authors 15% decreased word error rate and enhanced performance. The model uses spatiotemporal CNNs, RNNs, and the connectionist temporal classification (CTC) loss (Graves et al. [99]) and operates at the character level. The public sentence-level dataset GRID Corpus, published by Cooke et al. [100], was used for experiments. Another model designed for improved speed and accuracy is LipType [90]. In this work, poor light conditions are taken into consideration. As a first step, a spatial-temporal feature extraction technique was applied, which includes a correction for facial landmarks using Kalman filtering, 3D-CNN, and 2D SE-ResNet. Following that, Bidirectional Gated Recurrent Neural Network with CTC was used.

Dataset’s every class has approximately equal data number (see Table 2 for an exact size.) It is essential to mention that the dataset instance size is not equal to the version taken from Berkol et al. (2022) since we applied some necessary preprocessing steps to solve the lip reading problem with DL algorithms. The steps are explained in the following sections in detail. Also, we have more data examples for some classes since we added some noisy examples from our local data storage. Data will be updated as a new version. Additionally, we observe that the data

sequence length for each data sample depends on the length of the words and phrases. It can be concluded that the word or phrase length and frame number are highly correlated. Also, since the speakers are collected from a wide range of people, the dataset's classes are right-skewed, such as "merhaba" and "selam" which shows the speaker's speech speed differs.

Table 8.3. Size of the each class in the dataset.

Class	Number
günaydın	234
selam	235
merhaba	270
hoş geldiniz	230
özür dilerim	184

8.2.1. Applying classic CNN architecture for lip reading

The first model is the CNN model (see Figure 15). As it can be seen from Figure 15, two Conv3D layers with 96 filters and maxpooling3D layers are used as feature extraction layers. In the Conv3D layers, filters are applied with the size of (3, 3, 3), and strides are 1. Maxpooling3D layers applied with pooling size (2, 2, 2) and stride is 2. After Conv3D and maxpooling3D layers, flatten layer is applied. After that, two dense layers with 72 neurons were followed by a dropout layer with a probability of 0.4. Lastly, an output layer with 6 neurons is applied. Relu activation function is used in all layers except the output layer. In the output layer, the softmax activation function is used since we perform a classification problem with 6 classes. The other hyperparameters are as follows: the learning rate is 0.0002, the optimizer is Adam, and the loss function is categorical cross-entropy. Training is performed with early stopping monitoring validation accuracy, and patience is 4.

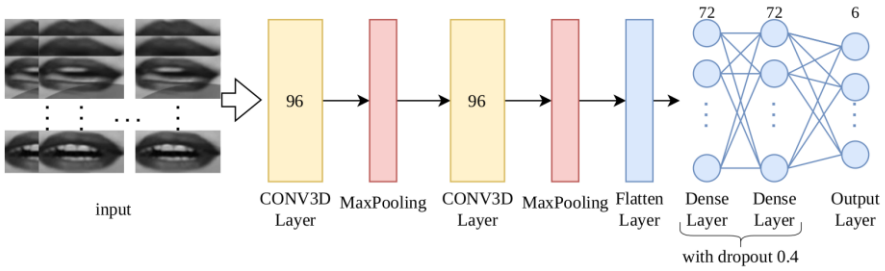

Figure 8.9. CNN model architecture

8.2.2. Applying LSTM model architecture for lip reading

The second model is the LSTM model (see Figure 16). The LSTM model is performed with two LSTM layers with 32 neurons and 0.5 dropout probability. Following that flatten layer is applied. The next layers are two dense layers with 64 neurons and 0.5 dropout probability. As an output layer, a dense layer with 6 neurons was applied. Except for the output layer, which uses the softmax function, the relu function is used in the fully connected layers. Other hyperparameters are as follows: the learning rate is 0.0002, the optimizer is Adam, and the loss function is categorical cross-entropy. Training is performed with early stopping monitoring validation accuracy, and patience is 5.

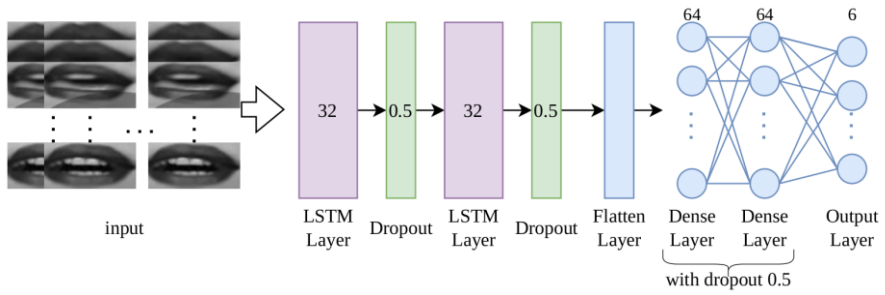

Figure 8.10. LSTM model architecture

8.2.3. Applying BGRU Model Architecture for lip reading

The last model is the BGRU model (see Figure 17). This model contains much fewer layers than the others. It uses a bidirectional GRU layer with 72 units and 0.2 dropout probability. Then, the flatten layer and dense layer with 64 neurons and 0.25 dropout probability. The last layer is again a dense layer with 6 neurons. As applied to the other models, the relu function is used in the hidden layer, and the softmax function is used in the output layer. The

other hyperparameters are as follows: the learning rate is 0.0001, the optimizer is Adam, and the loss is categorical cross-entropy. Similarly, the BGRU model is trained with early stopping monitoring validation accuracy, and patience is 3.

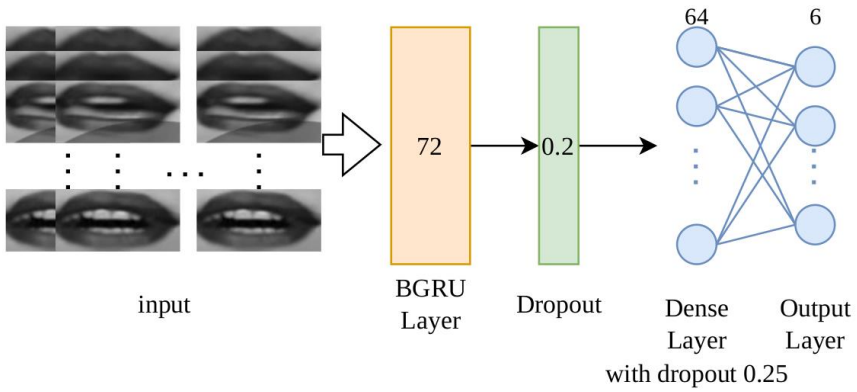

Figure 8.11. BGRU model architecture

The hyperparameter values are explained in detail in Table 2.

Table 8.4. Hyperparameters used in models. CCE: Categorical Cross Entropy.

Hyperparameter Name	CNN	LSTM	BGRU
learning rate	0.0002	0.0002	0.0001
optimizer	Adam	Adam	Adam
loss	CCE	CCE	CCE
hidden layer dropout	0.4	0.5	0.25
hidden layer neurons	72	64	64
hidden layer size	2	2	1
feature extraction layer	2	2	1
filter (CNN) /unit (LSTM, BGRU) s	96	32	72
feature extraction dropout prob.	-	0.5	0.2
activation function	ReLU	ReLU	ReLU
pooling size	2	-	-
patience	4	5	3

8.2.4. Comparative Results

The model architectures were explained in detail in the previous section. Experiments are run on an NVIDIA Tesla T4 graphics card. The dataset is divided into three parts: train, validation, and test sets with percentages of 70%, 15%, and 15%, respectively. The training set contains 3892 sets of examples, while validation and test sets contain 834 sets of examples. The training epochs are different since each model is trained with early stopping to prevent the model from overfitting. The CNN model's epoch size is 62, LSTM's epoch size is 58, and BGRU's epoch is 29. The accuracy results and training times for each model are in Table 3. The accuracy scores are very close to each other, unlike the training time. LSTM and BGRU models' accuracy scores are the same as the sixth decimal, 0.7781. CNN, which is 0.7649, performed the worst among the three models. In this case, training time helps decide the models' performance. The BGRU model is the fastest, approximately at 216 seconds, and the CNN model is the slowest, approximately at 863 seconds.

Table 8.5. Model accuracy and their training time results.

Model	Accuracy	Training time (secs)
CNN	76.49%	862.84
LSTM	77.81%	389.30
BGRU	77.81%	215.59

Additionally, we evaluated each model by confusion matrix (see Figures 18,19,20). Since the accuracy scores are almost the same, we observed that the confusion matrices of the LSTM and BGRU models differ. Phrases and words performed well among themselves for the three models. Moreover, we evaluated the precision, recall, and f1 scores for each class trained with the three models (see Table 4). As it can be seen from Table 4, there is no strict way to draw a conclusion about which model is more accurate. For example, for classes "hoş geldiniz" and "selam" CNN's precision scores are higher than others, or for classes "teşekkür ederim" and "merhaba" LSTM's precision scores are higher than the other two. However, we can observe that for some metrics and models, there is a considerably high difference between results.

Table 8.6. Comparision of Precision, recall, and f1 scores of models.

Words	Size	Model	Precision	Recall	F1 score
hoş geldiniz	153	CNN	0.6702	0.8366	0.7442
		LSTM	0.6089	0.8954	0.7249
		BGRU	0.6079	0.9020	0.7263
özür dilerim	105	CNN	0.6600	0.6286	0.6439
		LSTM	0.8594	0.5238	0.6509
		BGRU	0.8514	0.6000	0.7039
teşekkür ederim	139	CNN	0.8519	0.8273	0.8394
		LSTM	0.8264	0.8561	0.8410
		BGRU	0.8561	0.8129	0.8339
merhaba	167	CNN	0.8696	0.7186	0.7869
		LSTM	0.8639	0.7605	0.8089
		BGRU	0.8872	0.7066	0.7867
selam	141	CNN	0.8718	0.7234	0.790
		LSTM	0.8382	0.8085	0.8231
		BGRU	0.8014	0.8298	0.8153
günaydın	129	CNN	0.6993	0.8295	0.7589
		LSTM	0.8220	0.7519	0.7854
		BGRU	0.8197	0.7752	0.7968

For instance, "özür dilerim" class's precision score is much lower for the CNN model. On the other hand, "günaydın" class's recall score is much higher for the CNN model. For f1 score values, there are no such significant differences. To be more specific, the highest precision score, 0.88%, was obtained for "merhaba" with the BGRU model; similarly, the highest recall score, 90%, was obtained with the BGRU model on "hoş geldiniz", and the highest f1 score, 0.85%, was obtained with the LSTM model on "teşekkür ederim". If we consider the classes separately, we can conclude them as follows. Firstly, the phrases are evaluated. In the "hoş geldiniz" class, although the CNN model is the best in precision and f1 score, the recall score of the BGRU model is the highest among them. In the "özür dilerim" class, the LSTM model's precision is the best among all the models and metrics. CNN model is good at recall, and the BGRU model is good at the f1 score. The scores in the "teşekkür ederim" class are close. BGRU is the best in precision, and LSTM is the best for recall and f1 scores. Lastly, words are evaluated. In "merhaba" class, similar results with "teşekkür ederim" occur. BGRU is the best in terms of precision, and LSTM is the best for recall and f1 scores. In the "selam" class, the precision score

is the best with CNN, the recall score is the best with BGRU, and the f1 score is the best with LSTM. In the "günaydın" class, the precision score is the best with LSTM, the recall score is the best with CNN, and the f1 score is the best with BGRU.

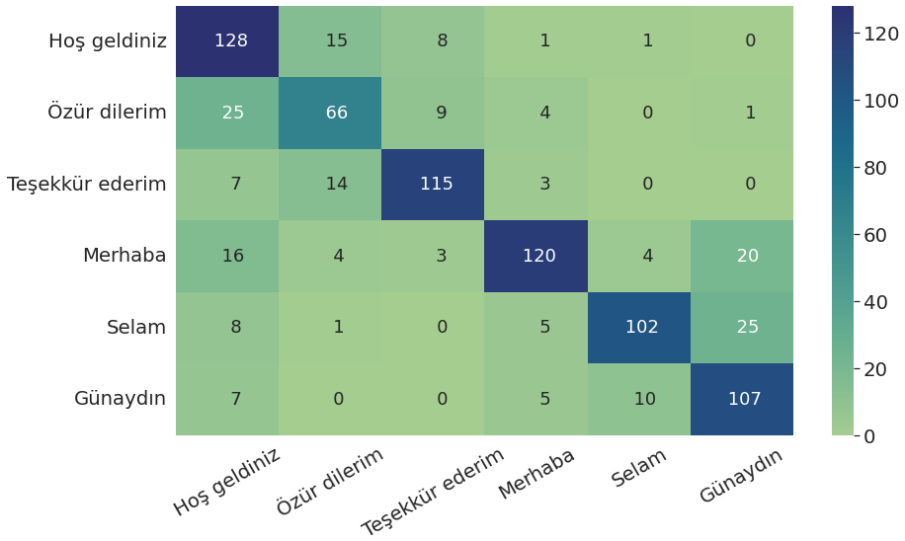

Figure 8.12. CNN model confusion matrix

Figure 8.13. LSTM model confusion matrix

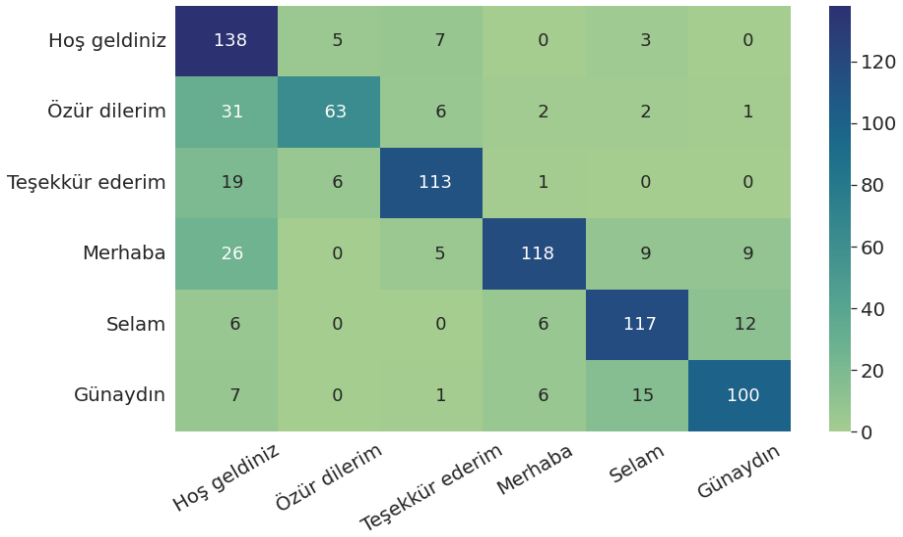

Figure 8.14. BGRU model confusion matrix

Appling data augmentation techniques such as horizontal flip, inverting pixel values, and sigmoid contrast techniques in order to enrich and diversify the data set. Additionally, we showed that different solution approaches, such as sequential and feature extraction techniques, can be used in the only-visual dataset. According to our experiments, recurrent-based models LSTM and BGRU proved their efficiency against the convolutional-based feature extraction technique CNN in terms of accuracy and training time. Hence, BGRU model is the most efficient when it is evaluated in terms of train time and overall classification results.

8.2.5. Dilated CNN Model

While creating the model 4, we built a Dilated CNN structure inspired by the temporal convolutional neural networks architecture Dilated CNN provides a more refined image by filtering some areas on the imageIn this architecture, we incorporated five consecutive dilated blocks. Each block consisted of spatial dropout and convolutional layers with dilation rates of 1, 2, 4, 8, and 1, respectively. By applying the dilation operation to the image, the input vector could be scanned in a broader and more efficient manner. Furthermore, since pixels in close proximity tend to have similar meanings, employing a more localized operation such as max pooling or dropout can be more effective than standard dropout. The utilization of dilated convolution and spatial dropout also serves as a means to prevent model overfitting. In the Add layers, the spatial dropout output is combined on the convolution layer. Final layer, the softmax

layer returns a score for three words and three phrases in Turkish. Also differences between Dilated CNN and Classical CNN shows below;

Table 8.7. Dilated CNN vs CNN

Differences	**Dilated CNN**	**CNN**
Convolution Operation	Convolution with dilation factor	Pixel-wise convolution
Receptive Field	Larger receptive field	Limited receptive field
Parameter Count	Fewer parameters	More parameters
Hierarchical Features	Better capture of hierarchical features	Good capture of hierarchical features

During the training process, hyperparameter tuning was conducted by experimenting with various values. The experimental studies involved exploring different values for the filter size of Dilated CNN layers, dilation rate, learning rate, input dimension, and the number of frames included in the training on the lip images within each sample. Early stopping was implemented to halt training if there was no improvement in the validation loss value. The dataset, consisting of 1,390 samples, was divided into 70% for training, 15% for testing, and 15% for validation to train and evaluate the model. The training was performed using parallel computation on a machine equipped with an NVIDIA GeForce GTX 1650 Ti graphics card with 4GB memory. Due to hardware limitations, a batch size larger than 4 could not be used for training. In comparison to our previous work, which took approximately 48 seconds for training using the CNN algorithm on the same machine, the training time increased to approximately 2 hours and 26 minutes for the more complex, multi-layered dilated CNN model.

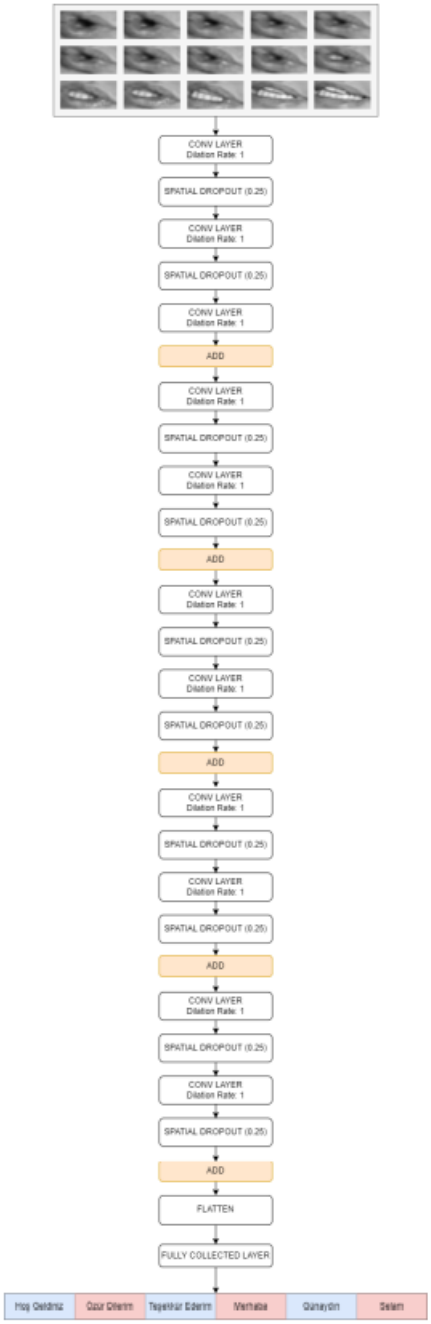

Figure 8.15. Dilated CNN model architecture

Table 8.8. Data train-validation-test split.

Train	Validation	Test	Total
973	208	209	1390

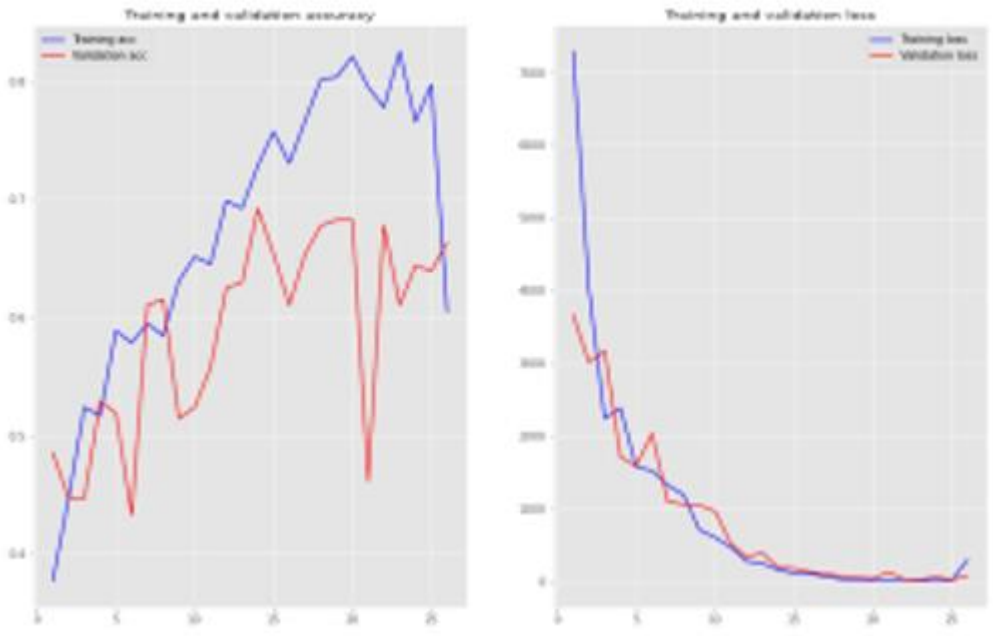

Figure 8.16. Dilated CNN training and validation loss and accuracy

Dilated CNNs have emerged as an alternative to traditional CNNs, especially in the field of image processing and segmentation. They excel at capturing a wider context and can achieve good performance with fewer parameters. However, both methods can be used in different contexts depending on the requirements of the task.

The trained model, which has early stopping strategy, is obtained at the end of 26 epochs for Dilated CNN. The validation/training accuracy and loss graph is obtained from the training process which is stopped automatically. As can be seen from the Fig.24 , if the training continues, the learning will continue, but since there will be no change in the validation loss value, it may cause the model to overfitting. We have evaluated and compared Dilated CNN model using accuracy, recall, precision and f1-score metrics to not ignore data diversity. The test accuracy we have obtained as a result of lip-reading studies for six words is 72% for Dilated CNN. In general, when we compare it with our previous work, CNN, it is seen that the standard CNN algorithm works better in terms of both time and performance. However, for some words, the detection performance is better compared to the overall accuracy, while for some words this score is lower. This is because the dataset from different youtube videos is diverse. We have tested Dilated CNN models with total number of 209 samples. In addition to diversity of data,

some words have more lip images than others. The Fig.25 show that the predictions made for each word and the words that resulted in incorrect predictions. When we look at the density in the diagonal, it is seen that there is mostly good performance for each word. Focuses on the words "merhaba" and "selam" which are incorrect predictions for the word "günaydın". In cases where "özür dilerim" and "teşekkür ederim" are guessed incorrectly, it should actually be "hoş geldiniz". If the interpretation is made according to these two situations, it can be said that the words and expressions are a prediction confusion in themselves.

In this we tried to develop a model which fits to real world. Although the dataset is challenging in both preprocess and training, we achieved remarkably good result in multi-class classification problem.

Table 8.9. Model Results for Dilated CNN

Words	Accuracy	Precision	Recall	F1-Score	Size
Hoş geldiniz (Welcome)	-	0.54	0.84	0.66	25
Özür dilerim (Sorry)	-	0.88	0.66	0.75	32
Teşekkür ederim (Thank you)	-	0.81	0.75	0.78	40
Merhaba (Hello)	-	0.70	0.70	0.70	33
Selam (Hi)	-	0.71	0.88	0.79	40
Günaydın (G.Morning)	-	0.78	0.54	0.64	39
All words	0.72	0.74	0.73	0.72	209

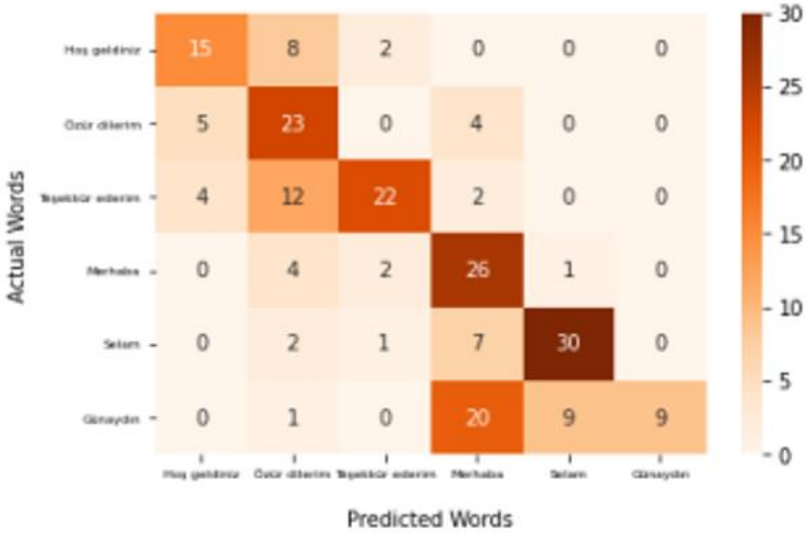

Figure 8.17. Confusion Matrix for Dilated CNN

8.2.6. Recommended CNN model

The proposed CNN architecture is two, for the concatenated frame lip images as a result of hyperparameter tuning, and for the lip images trained using discrete frames.

While collecting the data, it was tried to balance as much as possible with an equal number of samples for each class (see Table 8.10.).

Table 8.10. Data Distribution of Classes

Classes	Number of Samples
afiyet olsun	235
başla	235
bitir	244
görüşmek üzere	224
günaydın	232
hoş geldiniz	226
merhaba	268
özür dilerim	209
selam	235
teşekkür ederim	237

Since the dataset contains both single-word and 2-word classes, the pronunciation duration of the phrases varies. For example, since the phrases “teşekkür ederim” and “özür dilerim” are

longer, their pronunciation durations and the number of frames they occupy in the dataset are more than the word “selam”. While the frame count of the word “özür dilerim” exceeds 30, the frame count of the word “selam” does not exceed 15. It is critical to consider this distribution to make a balanced representation when classifying.

8.2.7. CNN model with discrete frame mouths input

As we mentioned in the Section Lip Representation, discrete frame lip images are given to the input layer as a sequence. The architecture includes two convolution and two max-pooling layers. Convolution layers use ReLU as an activation function, the filter sizes are 128, and the stride used in filters is 1 with no padding. Max-pooling layers pool sizes are 3x3x3 with the stride of 2. Flatten layer follows these four layers and architecture continues with fully connected layers with dropout.

The input vector consists of 15 images with a fixed size of 50x50. Random 128 filters are applied to these images in the convolution layer without padding and with a stride of 1 step. After the convolution process, an output of 13x48x48x128 is produced. Since there is a 3x3 pool size in the output of the max pooling layer following the convolution, it outputs as 6x24x24x128. After applying the conv3d, max-pooling, and flatten layers, respectively, a 15488 dimensional vector is obtained. Two fully connected layers with ReLU activation function and 0.5 ratio dropout layers used to avoid overfitting, especially in CNN models are implemented. Finally, since a multi-class classification problem is studied, the architecture is finalized with a fully connected layer that produces 10-dimensional vector output with the softmax activation function. In the output, probabilities are produced for 10 classes in the form of “afiyet olsun”, “başla”, “bitir”, “görüşmek üzere”, “günaydın”, “hoş geldiniz”, “merhaba”, “özür dilerim”, “selam”, and “teşekkür ederim”.

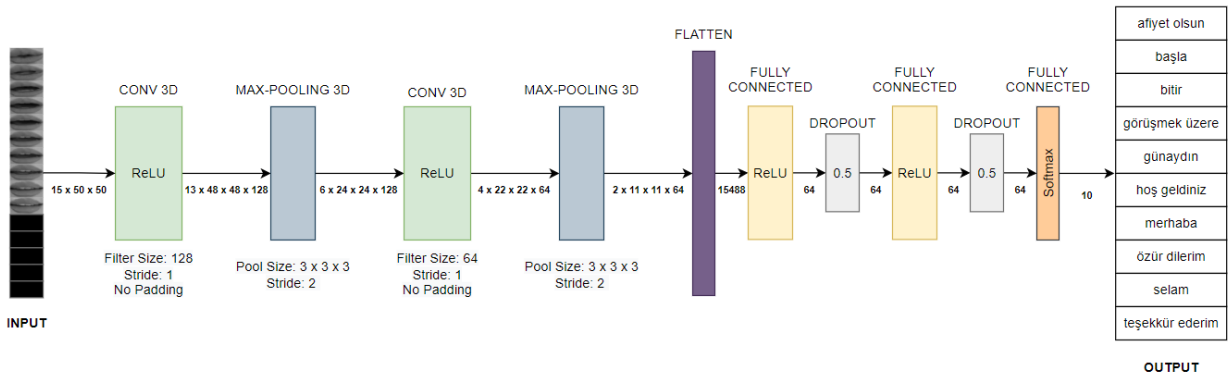

Figure 8.18. CNN Model using discrete frame represented mouths

8.2.8. CNN model with concatenated frame mouth input

In this approach where lips are combined, 15 images are concatenated to form a single image input, unlike the case of discrete frame mouths as input. Therefore, it is quite convenient in terms of computational cost. Experiments were conducted using a shallower series of convolution layers compared to the previous CNN model, since a single image represents a sequence of images, reducing data complexity. It is sent to the convolution layer using a 50x50 image as input. Experiments were conducted using a shallower series of convolution layers compared to the previous CNN model, since a single image represents a sequence of images, reducing data complexity. It is sent to the convolution layer using a 50x50 image as input. Then the Flatten layer's input is 24x24x16 since the pool size is 2x2. Unlike the architecture in the approach where the lips are given separately, there is 1 fully connected layer and dropout after the Flatten layer, which has 9216 dimensional vector output. Finally, an output vector with 10 classes is produced.

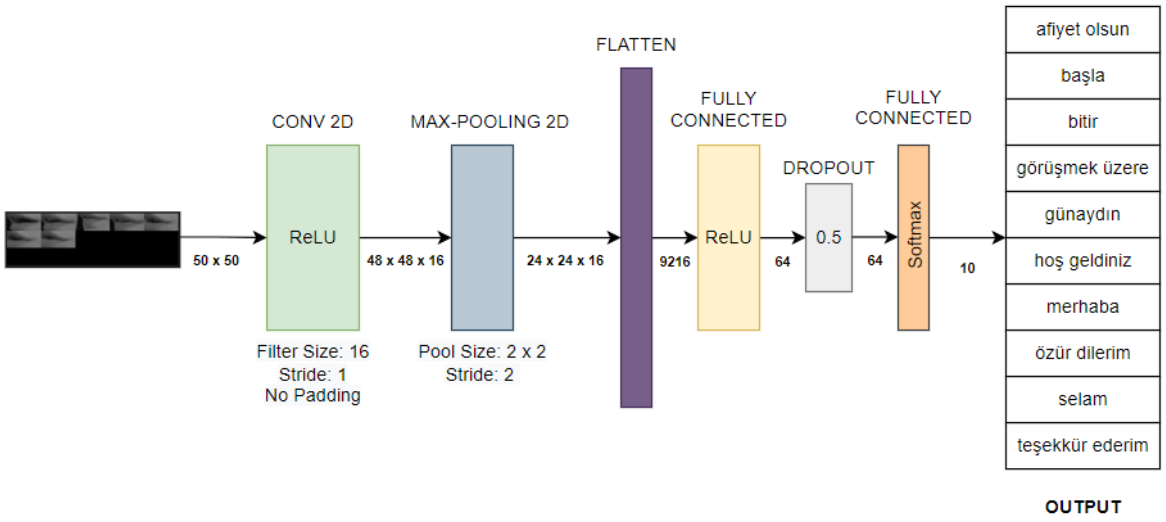

Figure 8.19. CNN Model using concatenated frame represented mouths

8.2.9. Training

In the training process, experiments were carried out on different hyperparameters for studies on two different approaches to training the lips separately and combining them. Mlflow, a Python library developed to manage the machine learning lifecycle, was used to evaluate the results of the experiments and make hyperparameter tuning. In Table 2, it is seen that the hyperparameter results for both approaches.

Different hyperparameters have been applied for the cases where the lips are joined and separate. Since the model capacity and complexity of the two approaches are different, parameters such as learning rate, batch size, and number of epochs varied. Also, Categorical cross entropy is an information measure used to compare predictions with true labels in a classification problem.

This method quantifies the difference between the predicted probability distributions and the true labels, thereby measuring the accuracy of the model.

A higher cross entropy value indicates that the predictions are further away from the true labels, while a lower value signifies a better match.

Table 8.11. CNN model training parameters

Parameters	Discrete Frame Mouth	Concatenated Frame Mouth
Number of training samples	1606	1606
Number of validation samples	345	345
Number of test samples	344	344
Learning rate	0,0002	0,002
Batch size	32	16
Word length	15	15
Input dimension	50	50
Loss function	Categorical cross entropy	Categorical cross entropy
Optimizer	Adam	Adam
Total trainable parameters	1,220,938	590,698

8.2.10. Results

It is difficult to make an accurate assessment in studies where language is involved, such as lip reading, because there are different pronunciations and variations in the language. It is possible to make an evaluation, especially when there are many studies and data in the English language. However, there is no comparable word-level dataset in terms of our studies in Turkish.

In our studies, we basically aimed to develop a CNN architecture for the Turkish lip reading problem. All experiments based on CNN architecture run on GPU. NVIDIA 1650 Ti graphics card with 4GB memory. The improvements were made using the Python Keras library. In addition to these, Plotly and Seaborn libraries were used for visualization, and OpenCV libraries were used for image processing studies.

Experiments were performed with the number of samples in Table 3 in both of the representation steps, where the lips are joined and the lips are separated.

Table 8.12. Number of test samples of each class

Classes	Number of Samples
afiyet olsun	29
başla	35
bitir	46
görüşmek üzere	23
günaydın	32
hoş geldiniz	34
merhaba	43
özür dilerim	31
selam	37
teşekkür ederim	35

8.2.11. Training Results with Discrete Frame Lips

Looking at the training results, the accuracy and loss value changes for which the epoch number is determined using early stopping are shown in the Fig. 45.

The training process, which was stopped after the improvement in Loss value did not improve in 3 epochs, ended in 68 epochs. If the training continues further, there is no need to make further calculations as the model will be overfitting.

When the results of the predicted classes in the test data are examined, it is seen that the incorrectly determined classes are generally collected in the “afiyet olsun” class, see Fig. 10. Especially for instances of classes whose actual class is “başla”, “günaydın”, and “özür dilerim”, the wrong predictions concentrated on “afiyet olsun”. Mistakes made in the “afiyet olsun” class were generally made for 6 examples in the “teşekkür ederim” phrase. Contrary to these, there is no example of an incorrectly guessed “afiyet olsun” in the “hoş geldiniz” phrase.

Figure 8.20. Training and validation accuracy and loss per epoch with discrete frame lips

“hoş geldiniz”, “merhaba”, “selam”, and when looked at the “başla”, “bitir”, “özür dilerim” classes that follow them, it is seen that the precision scores are high, see Fig. 46,47. Thus, we can interpret that the majority of positive predictions for these classes are correct. In general, we see that the “afiyet olsun” class error rate is high based on the confusion matrix. There may not be a clear lip movement in the vocalization of these phrases in the dataset, or it may be interpreted as one of the more challenging expressions compared to Turkish grammar rules. Since f1-score is the harmonic mean of precision and recall metrics, it is generally seen as f1-score high when precision and recall are high at the same time, or low when f1-score is low at the same time such as “hoş geldiniz” and “günaydın”. Although there is no class imbalance in terms of the number of samples in this dataset, the prediction performances vary according to the classes, as there are situations that create diversity for each class, such as the differences in speakers, viewing angles, and light differences, just like real-life scenes.

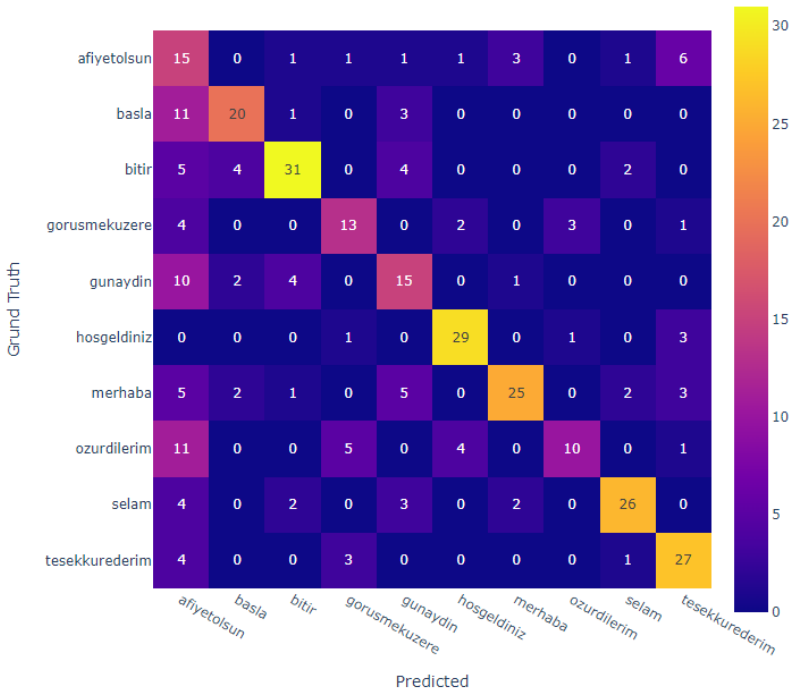

Figure 8.21. Confusion matrix of model trained with discrete frame lips

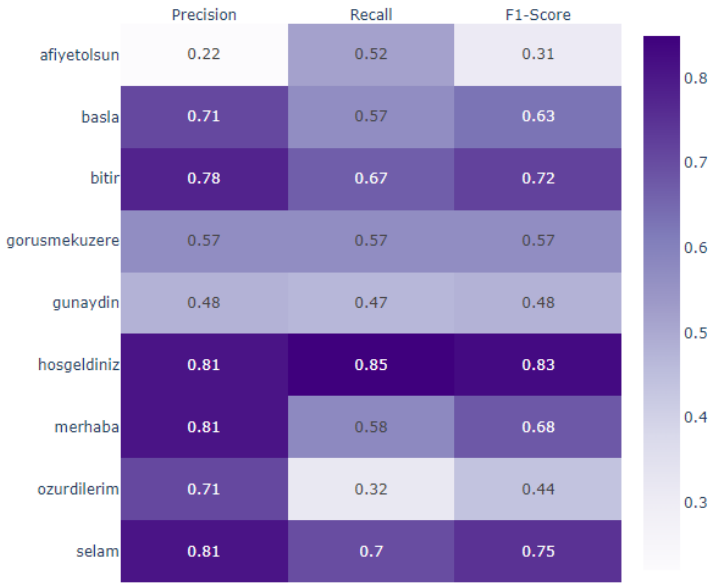

Figure 8.22. Classification report of model trained with discrete frame lips

8.2.13. Training results with concatenated frame lips

In the CNN model training performed using joined lips, the process stopped with early stopping ended at 42 epochs (see Fig. 48). In the last epochs, validation accuracy starts to decrease, while training accuracy increases. Therefore, if the training continues further, it will be inevitable to achieve a low test accuracy.

Similarly, as in the dataset with split lip images, wrong predictions for many classes such as “bitir”, “günaydın”, “merhaba” and “özür dilerim” in the results of combined lip images were collected in the “afiyet olsun” class, see Fig. 49. Apart from that, we can see that the estimations are generally high in the “başla”, “görüşmek üzere”, “hoş geldiniz”, “selam” and “teşekkür ederim” classes and do not predominantly confused with other classes.

As seen in the confusion matrix, it is observed in the classification report graph (Fig. 14) that the precision, recall and f1-score values of the “afiyet olsun” class are low. To interpret the accuracy percentages of other classes, more balanced results are seen compared to training using split lips.

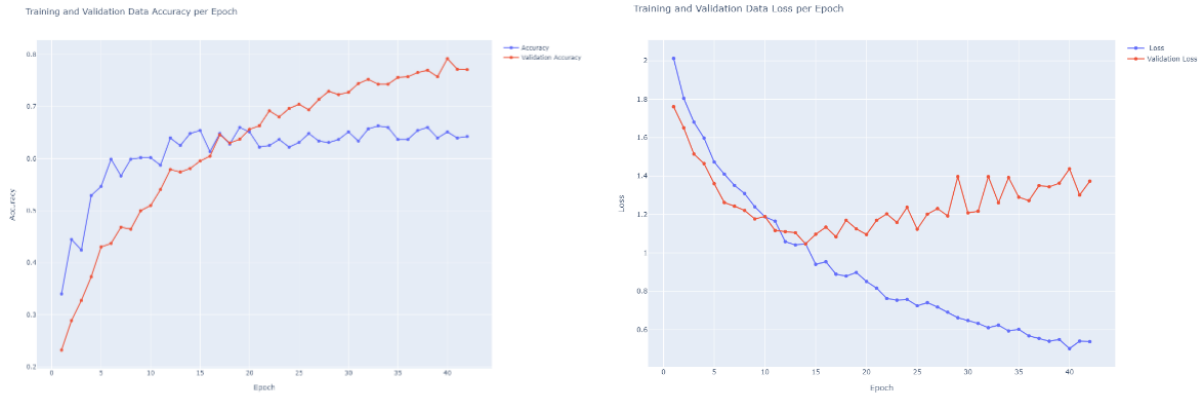

Figure 8.23. Training and validation accuracy and loss per epoch with concatenated frame lips

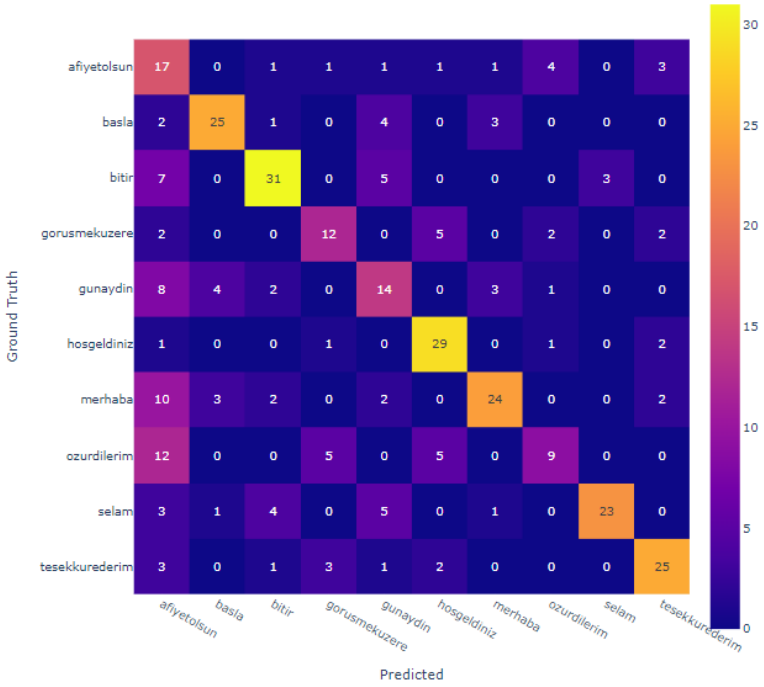

Figure 8.24. Confusion Matrix of Model Trained with Concatenated Frame Lips

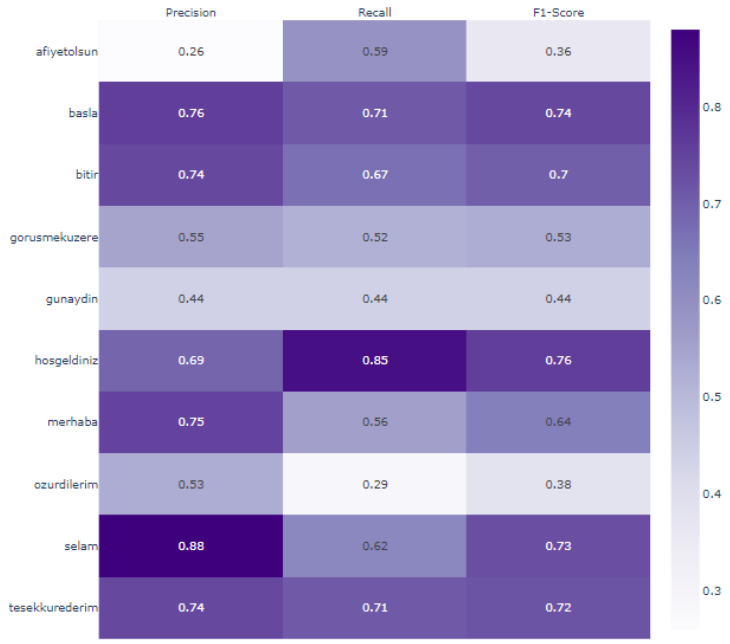

Figure 8.25. Classification Report of Model Trained with Concatenated Lips

8.2.14. Comparision for concatenated frame lips results and discrete frame lips results

When the total experimental results are compared, the accuracy is 90.6% for concatenated frame lips and 91.7% for discrete frame lips. Again, the times are 18 seconds and 8 minutes, respectively. Since the training time of the discrete frame lips is long, it can be considered more burdensome in terms of computational cost, but it can be preferred in terms of performance because of its higher accuracy. In terms of image representation, 15 images of 50x50 size are used in one of the inputs, while 1 image of 60x200 size is used in the other. In a situation where simultaneous estimation is required, the use of representation using joined lips would be more appropriate, but for problems where accurate detection is important, the use of the CNN model using split lips is appropriate.

Compared to similar studies, the data contents used in terms of the dataset are quite challenging. In this novel dataset, faces are not viewed from the front, some images are very dark while others are quite bright, and at times it is not possible to accurately detect the face because the background is too mixed.

Table 8.13. Accuracy and training time of two CNN models

	Accuracy	**Training Time**
Concatenated Frame Lips	90.6%	18 seconds
Discrete Frame Lips	91.7%	480 seconds

9. CONCLUSION AND SUGGESTIONS

In this thesis, a CNN model is proposed for a new Turkish dataset. It also compares accuracy and computational cost with two different input representations. In the first of these, sequence lip images form the input of the model separately, while in the other, the lips are combined to form a single image. In terms of performance, split lips look better, but combined lips perform better in terms of time cost. In addition, the Turkish dataset collected from natural Youtube images is also challenging as it is closer to real-world images compared to other studies. The images collected in the studies in the literature were obtained with a fixed background and a fixed human pose by establishing a controlled environment. There is a known dataset that can be evaluated for Turkish, although it has more data, it was also collected in a controlled environment. Automatic lip-reading over natural videos is also of great importance in terms of automatic captioning for hearing-impaired people. In this study, a CNN model is proposed by performing lip reading from natural video images. Since the natural language and lip reading studies in Ural-Altaic languages are shallow, we have contributed with a unique study.

On the image sequences, the lip-reading problem was handled by making multiclass classification with frequently used greeting words in Turkish. Images consisting of frame sizes in different numbers obtained from the video are used. These classified images are more challenging for the visibility of the lips than datasets obtained in a controlled environment.

A benchmarking was made by classifying the lip images collected in the natural environment with CNN and three other classification algorithms, which are the approaches that are frequently used in deep learning. The performance of the newly collected dataset was evaluated on the basic approaches. This evaluation aimed to assess the effectiveness of the CNN model compared to other commonly used classification algorithms in the context of lip image classification. The results of the benchmarking provided insights into the performance and suitability of different algorithms for lip image classification tasks in natural environments.

The dataset used in classification studies consists of frequently encountered greeting expressions in Turkish. These words were collected from natural videos, ensuring that they were obtained without any intervention such as clipping or elimination. This dataset, known as the Turkish lip reading dataset, is publicly available for research purposes. One notable distinction

between this dataset and many other lip reading datasets is that it contains solely image data, without accompanying audio. This characteristic makes it particularly suitable for scenarios where audio data is unavailable or absent in the environment. Researchers and practitioners can utilize this dataset to explore and develop lip reading systems specifically designed for Turkish greeting expressions. The availability of such a dataset contributes to advancing the field of lip reading and further supports the development of robust models for speech recognition and understanding.

Another significant contribution of this study is the evaluation of discrete and concatenated representations of the collected lip data within a CNN architecture that exhibits promising performance. In addition to the classification task, particular attention has been given to accurately distinguish faces in the images and subsequently identify the lip region. This comprehensive evaluation aims to explore the effectiveness of different representations of lip data in the context of a CNN architecture.

The collected lip data is represented using two distinct approaches: discrete representation and concatenated representation. The discrete representation involves treating each lip image as an individual data point, while the concatenated representation involves combining multiple lip images to form a single input. By evaluating the performance of these representations, the study seeks to determine which approach yields better results in terms of accuracy and efficiency.

Furthermore, an emphasis has been placed on the approaches employed to accurately distinguish faces within the images. This involves employing techniques such as face detection and facial landmark localization to precisely identify the lip region of interest. The effectiveness of these face detection and lip localization methods is also assessed in the study, contributing to the overall understanding of the lip image classification process.

Through this comprehensive examination, the study aims to provide insights into the optimal representation of lip data and the effectiveness of face detection and lip localization techniques in the context of a CNN architecture. The findings will further advance the field of lip image classification and contribute to the development of more accurate and efficient lip reading systems

Furthermore, the study places a significant focus on achieving accurate identification of faces within the images. Recognizing the importance of precise face detection as a crucial initial step in the subsequent lip image identification process, the research explores various approaches

and techniques to ensure optimal face detection performance. The evaluation encompasses the assessment of different algorithms, methodologies, and pre-processing steps employed to achieve reliable and accurate face detection.

Following successful face detection, the study's emphasis shifts to the identification and extraction of lip images. Once the faces are accurately detected, the research investigates different approaches employed to discern and extract the lip region from the overall facial image. Various techniques, such as image segmentation, feature extraction, and pattern recognition, are thoroughly examined and their effectiveness is evaluated in terms of accurately isolating the lip region.

For image segmentation, different algorithms are explored to partition the facial image and separate the lip region from the rest of the face. These algorithms aim to precisely identify the boundaries of the lips and separate them from other facial components. Feature extraction techniques are then applied to extract relevant visual characteristics and discriminative information from the lip region. These features play a crucial role in distinguishing between different lip shapes, movements, and articulations. Additionally, pattern recognition methods are employed to recognize and classify the extracted lip images into relevant categories or classes.

By investigating and evaluating these approaches for both face detection and lip image identification, the study aims to contribute to the advancement of accurate and reliable lip reading systems. The findings will provide valuable insights into the effectiveness of different techniques, algorithms, and pre-processing steps in the context of face and lip analysis, ultimately enhancing the overall performance of lip image classification and recognition tasks.

By conducting these comprehensive evaluations and in-depth analyses, this study aims to make significant contributions to the advancement of lip image recognition techniques within the domain of computer vision. The findings and insights gained from this research have the potential to enhance the accuracy, reliability, and overall performance of lip reading systems, biometric authentication applications, and other related fields where lip image analysis plays a crucial role.

The outcomes of this study can pave the way for improved lip image classification and recognition algorithms, leading to more robust and efficient systems. The research outcomes can also guide the development of novel approaches and methodologies for face detection, lip

region identification, and subsequent analysis. These advancements are particularly valuable in areas such as human-computer interaction, assistive technologies for speech-impaired individuals, and biometric security systems where accurate lip image analysis can provide valuable information for identification and authentication purposes.

Additionally, the insights obtained from this study can inform the design and optimization of lip reading systems for diverse applications, including transcription services, automatic speech recognition, and audiovisual synchronization. The advancements in lip image recognition can contribute to the development of inclusive and accessible technologies, enabling effective communication and interaction for individuals with hearing impairments or in noisy environments.

Overall, the findings and contributions of this study have the potential to advance the field of lip image recognition, benefiting a wide range of domains that rely on accurate and efficient analysis of lip-related visual information. Through this research, advancements in computer vision techniques can be harnessed to unlock new possibilities and applications, ultimately enhancing communication, security, and accessibility in various real-world scenarios.

Some suggestions for further Works;

Utilizing Depth Information: To enhance lip movements with more information, you can consider integrating data obtained from depth cameras into your system. By incorporating both 2D image data and 3D depth information, a more precise lip reading system can be developed.

Multilingual Support: Although your current focus is on the Turkish language, providing multilingual support for your system can cater to a broader user base. Collecting data for different languages and investigating the language-dependent characteristics of lip movements can present future opportunities for research.

Real-time Application: Evaluating the performance of your lip reading system in real-time scenarios is crucial. Developing a system capable of analyzing live video streams in real-time would be a significant step towards real-world applications by using Tiny ML & Edge AI.

LSTM variations: Explore different variations of LSTM beyond the traditional LSTM architecture. For instance, you can investigate the usage of peephole connections, gated units at the cell level, or stacked LSTM structures.

Transfer learning: Investigate the potential usage of an LSTM model trained on another language (e.g., English) to recognize Turkish lip movements. Transfer learning can be beneficial when working with limited data.

REFERENCES

[1] B. Bir, A. A. Fenton, and U. Rutishauser, "Spiking neural networks for BCI applications," *Trends in Neurosciences*, vol. 43, no. 6, pp. 443-454, 2020.

[2] H. Höge, V. Dellwo, and R. Wright, *Lip Reading: A Scientific Approach to a Hidden Mode of Communication,* New York: IGI Global, 2019.

[3] W. C. Liew and S. Wang, *Visual Speech Recognition: Lip Segmentation and Mapping: Lip Segmentation and Mapping,* New York: IGI Global, 2009.

[4] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson, "The mcgurk effect in infancy: looking to the future," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 42, no. 10, pp. 1397-1411, 2016.

[5] E. T. Auer and L. E. Bernstein, "Speechreading and the structure of the lexicon: computational and behavioral evidence," *Frontiers in Psychology*, vol. 8, p. 1282, 2017.

[6] L. Johanson and É. Á. Csató, *The Turkic Languages*, New York: Routledge, 1998.

[7] J. Janhunen, *The Mongolic Languages,* New York: Routledge, 2003.

[8] M. Robbeets, *The Oxford Guide to the Transeurasian Languages*, New York: Oxford University Press, 2017.

[9] D. Sinor, *The Cambridge History of Early Inner Asia,* London: Cambridge University Press, 1990.

[10] N. Poppe, *Introduction to Altaic Linguistics*, Wiesbaden: Otto Harrassowitz, 1965.

[11] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212-215, 1954.

[12] Q. Summerfield, "Lipreading and audiovisual speech perception," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 316, no. 1177, pp. 303-315, 1987.

[13] R. Campbell, "The Processing of audio-visual speech: empirical and neural bases," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 1001-1010, 2008.

[14] N. P. Erber, "Interaction of auditory and visual information in speech perception," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 423-425, 1969.

[15] D. W. Massaro and M. M. Cohen, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, New York: The MIT Press, 1995.

[16] L. E. Bernstein, E. T. Auer, and J. K. Moore, "Audiovisual speech binding: convergence or association?" *Cognition*, vol. 92, no. 3, pp. 229-261, 2004.

[17] Q. Summerfield, "Lipreading and audiovisual speech perception," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 71-78, 1992.

[18] S. Petridis, M. Pantic, and O. Mayora, "Audiovisual classification of vocal outbursts in spontaneous human interaction," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 139-151, 2009.

[19] G. Potamianos and H. P. Graf, "Visual speech information processing for automatic speech recognition," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1396-1420, 2002.

[20] Küblbeck and A. Zisserman, "Face recognition with support vector machines: global versus component-based approach," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, San Diego, 2005, pp. 688-695.

[21] M. Gurban and J. P. Thiran, "Audio-Visual speech recognition using continuous hidden markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, 2006, pp. 221-224.

[22] G. Gnecco and S. Mozziconacci, "Visual speech recognition using particle filters," *Pattern Recognition Letters*, vol. 28, no. 15, pp. 1984-1992, 2007.

[23] Y. Lee, S. W. Park, and H. C. Shin, "Visual Speech recognition using lip information extracted by active appearance model," *Pattern Recognition Letters*, vol. 31, no. 15, pp. 2353-2360, 2010.

[24] P. Varga and J. P. Lewis, "Hidden markov model-based viseme classification for visual speech recognition,*" IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 74-83, 2002.

[25] U. Tariq, M. Ali, and M. K. Shahid, "Audio-visual speech recognition using hidden markov models," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 2, pp. 123-131, 2014.

[26] M. Ali, M. Liwicki, and A. Dengel, "An HMM-based approach for viseme classification," in *Proceedings of the 10th International Conference on Document Analysis and Recognition*, Barcelona, 2011, pp. 833-837.

[27] M. Hasegawa-Johnson and X. Huang, "Cross-Modal speaker adaptation for audio-visual speech recognition," *Computer Speech & Language*, vol. 19, no. 2, pp. 135-153, 2005.

[28] H. Han, S. Kang, and C. D. Yoo, "Multi-view visual speech recognition based on multi task learning," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, 2017, pp. 3983-3987.

[29] L. Ashok Kumar, D. Karthika Renuka, S. Lovelyn Rose, M. C. Shunmuga priya, I. Made Wartana, "Deep learning based assistive technology on audio visual speech recognition for hearing impaired," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 24-30, 2022.

[30] S. Petridis, M. Pantic, and P. Maja, "Deep facial expression recognition: A survey," *Image and Vision Computing*, vol. 66, pp. 115-129, 2018.

[31] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1713-1727, 2018.

[32] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1691-1695, 2017.

[33] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 2017, pp. 3444-3453.

[34] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016, pp. 3444-3453.

[35] T. Gergen and H. K. Ekenel, "Visually inspired deep neural networks for spoken word recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1871-1883, 2016.

[36] M. Wand, J. Koutník, J. Schmidhuber, and R. Memisevic, "Lipreading with long short-term memory," in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), Shanghai, 2016, pp. 6115-6119.

[37] S. Petridis, P. Ma, M. Pantic, and A. Nijholt, "Visual speech recognition using dynamic visemes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1410-1423, 2016.

[38] T. Gergen and H. K. Ekenel, "Deep neural networks for visually inspired speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, 2015, pp. 4545-4549.

[39] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, 2014, pp. 2590-2597.

[40] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Total capture: A 3D deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, 2014, pp. 832-839.

[41] Z. Zhou and R. Chellappa, "Context-aware deep convolutional neural networks for exploratory audio-visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Istanbul, Turkey, 2012, pp. 3506-3513.

[42] Shokouhi and C. Busso, "Visual speech recognition using deep neural networks," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2012, pp. 4177-4180.

[43] Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.

[44] Kaya, H. Özer, and G. Ercan, "Turkish visual speech recognition using deep convolutional neural networks," *Journal of Signal Processing Systems*, vol. 91, no. 5, pp. 569-579, 2019.

[45] Demirel and G. Ercan, "Turkish audio-visual speech recognition using deep neural networks," in *2018 IEEE 26th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, 2018, pp. 1-4.

[46] R. Kılıç and E. Şahin, "Turkish word level lipreading using convolutional neural networks," in *2020 IEEE 28th Signal Processing and Communications Applications Conference (SIU)*, Ankara, Turkey, 2020, pp. 1-4.

[47] M. A. Bilgin and E. Erzin, "Turkish sentence level lipreading using convolutional neural networks," in *2017 IEEE 25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, 15-18 May 2017, pp. 1-4.

[48] S. Göktürk, "Turkish phoneme recognition from lip motion using deep learning models," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, Gaziantep, Turkey, Oct. 5-8, 2020, pp. 1-4.

[49] Erol and M. Ersahin, "Turkish speaker verification based on lip movement using deep learning," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, Denizli, Turkey, Apr. 24-26, 2019, pp. 1-4.

[50] Gergen, M. Kolbæk, and T. B. Moeslund, "Deep lip reading using large-scale multi-task learning," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Istanbul, Turkey, May, 18-22, 2020, pp. 357-364.

[51] S. Bekesiene, R. Smaliukiene, and R. Vaicaitiene, "Using artificial neural networks in predicting the level of stress among military conscripts," *Mathematics*, vol. 9, no. 6, pp. 626, 2021.

[52] K. Zarzycki and M. Ławryńczuk, "LSTM and GRU neural networks as models of dynamical processes used in predictive control: a comparison of models developed for two chemical reactors," *Sensors*, vol. 21, no. 16, pp. 5625, 2021.

[53] Y. Chen, H. Liu, X. Wang, and Y. Gao, "LipNet: End-to-end sentence-level lipreading," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Salt Lake City, UT, USA, June. 18-22, 2018, pp. 3444-3453.

[54] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, June. 27-30, 2016, pp. 3444-3453.

[55] Gan, J. Li, and Y. Gong, "Convolutional sequence to sequence model for human dynamics," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, June. 27-30, 2016, pp. 3444-3453.

[56] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "multimodal deep learning," in *2011 28th International Conference on Machine Learning (ICML),* Bellevue, WA, USA, June. 28, 2011, pp. 689-696.

[57] J. S. Chung and A. Zisserman, "VoxCeleb2: deep speaker recognition," in *Interspeech,* Hyderabad, India, Sept. 2-6, 2018, pp. 1086-1090.

[58] S. Petridis, M. Pantic, and D. Coniam, "End-to-End visual speech recognition with LSTMs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* vol. 40, no. 9, pp. 2097-2109, 2018.

[59] T. Afouras, J. S. Chung, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* vol. 41, no. 5, pp. 1025-1038, 2019.

[60] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition using multiple streams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, AB, Canada, Apr. 15-20, 2018, pp. 2572-2576.

[61] W. Hu, Y. Zhang, and L. Li, "Study of the application of deep convolutional neural networks (cnns) in processing sensor data and biomedical images," *Sensors*, vol. 19, no. 16, pp. 3584, 2019.

[62] S. S. Roy, N. Rodrigues, and Y.-h. Taguchi, "Incremental dilations using CNN for brain tumor classification," *Appl. Sci.,* vol. 10, no. 14, pp. 4915, 2020.

[63] Ü. Atila and F. Sabaz, "Turkish lip-reading using Bi-LSTM and deep learning models," *Engineering Science and Technology, an International Journal*, vol. 35, pp. 101206, 2022.

[64] Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, pp. 198-213, 2002.

[65] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, pp. 1-1, 2019.

[66] "Lip Reading Sentences 2 (LRS2) Dataset," [Online]. [Accessed: 23-Sep-2022].

[67] "Lip Reading Sentences 3 (LRS3) Dataset," [Online]. [Accessed: 23-Sep-2022].

[68] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv*:1809.00496, 2018.

[69] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, May. 14-18, 2019, pp. 1-8.

[70] Egorov, V. Kostyumov, M. Konyk, and S. Kolesnikov, "LRWR: large-scale benchmark for lip reading in Russian language," *arXiv*:2109.06692, 2021.

[71] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Computer Vision and Image Understanding*, vol. 173, pp. 76-85, 2018.

[72] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, Oct. 8-16, 2016, pp. 285-301.

[73] Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, Slovenia, May. 4-8, 2015, pp. 1-5.

[74] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Computer Science*, vol. 163, pp. 400-409, 2019.

[75] P. Sujatha and M. R. Krishnan, "Lip feature extraction for visual speech recognition using hidden Markov model," in *Proceedings of the 2012 International Conference on Computing, Communication and Applications*, Dindigul, India, Feb. 22-24, 2012, pp. 1-5.

[76] R. Shashidhar and S. Patilkulkarni, "Visual speech recognition for a small-scale dataset using VGG16 convolutional neural network," *Multimedia Tools and Applications*, vol. 80, pp. 28941-28952, 2021.

[77] B. Xu and H. Wu, "Lip reading using multi-dilation temporal convolutional network," in *CONF-SPML Signal Process. Mach. Learn.*, vol. 3150, pp. 50-59, 2022.

[78] J. S. Chung and A. Zisserman, "Lip reading in the wild," *in Proceedings of the 13th Asian Conference on Computer Vision (ACCV)*, Taipei, Taiwan, Nov. 20-24, 2016, pp. 87-103.

[79] Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *arXiv*:2011.07557, 2020.

[80] Berkol, N. P. Akman, T. T. Sivri, and H. Erdem, "Lip reading multiclass classification by using dilated CNN with Turkish dataset," in *Proceedings of the 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET),* Prague, Czech Republic, July. 20-22, 2022, pp. 1-6.

[81] M. Cooke, J. Barker, S. Cunningham, and S. Xu, *The Grid Audio-Visual Speech Corpus (1.0)*, Geneva, Switzerland: Zenodo, 2006.

[82] Rekik, A. Ben-Hamadou, and W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," in *Proceedings of the 11th International Conference on Image Analysis and Recognition (ICIAR 2014),* Vilamoura, Portugal, Oct. 22-24, 2014, pp. 21-28.

[83] Desai, P. Agrawal, P. Parikh, and P. K. Soni, "Visual speech recognition," *International Journal of Engineering Research and Technology*, vol. 9, no. 4, pp. 601-605, 2020.

[84] YouTube [Online] (accessed on 17 October 2022).

[85] G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language, and Hearing Research*, vol. 11, no. 4, pp. 796-804, 1968.

[86] R. D. Easton and M. Basala, "Perceptual dominance during lipreading," *Perception and Psychophysics*, vol. 32, no. 6, pp. 562-570, 1982.

[87] S. Lesani, F. F. Ghazvini, and R. Dianat, "Mobile phone security using automatic lip reading," in *9th International Conference on e-Commerce in Developing Countries: With a focus on e-Business (ECDC),* Isfahan, Iran, 2015, pp. 1-5.

[88] S. Mathulaprangsan, C. Y. Wang, A. Z. K. Frisky, T. C. Tai, and J. C. Wang, "A survey of visual lip reading and lip-password verification," in *International Conference on Orange Technologies (ICOT),* Hong Kong, China, 2015, pp. 22-25.

[89] Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March. 20-25, 2016, pp. 4945-4949.

[90] J. T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, Apr. 19-24, 2015, pp. 4989-4993.

[91] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Shanghai, China, Dec.13-17, 2016, pp. 167-174.

[92] Hyunmin, C. M. Kang, B. Kim, J. Kim, C. C. Chung, and W. Choi, "Autonomous braking system via deep reinforcement learning," *ArXiv*, abs/1702.02302, 2017.

[93] Soltani, F. Eskandari, and S. Golestan, "Developing a gesture-based game for deaf/mute people using microsoft kinect," in *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, Palermo, Italy, Jul. 4-6, 2012, pp. 491-495.

[94] J. Tan, C. T. Nguyen, and X. Wang, "SilentTalk: Lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, Atlanta, GA, USA, May. 1-4, 2017, pp. 1-9.

[95] J. Tan, X. Wang, C. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1-18, 2018.

[96] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 447-460, 2019.

[97] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, Jul. 21-26, 2017, pp. 6447-6456.

[98] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2, pp. 1-9, 2007.

[99] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," *IEEE Access*, vol. 8, pp. 215516-215530, 2020.

[100] L. Pandey, A. S. Arif, "LipType: a silent speech recognizer augmented with an independent repair model," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21),* Yokohama, Japan, May. 8-12, 2021, pp. 1-19.

[101] M. Faisal, S. Manzoor, "Deep learning for lip reading using audio-visual information for urdu language," *ArXiv*, abs/1802.05521, 2018.

[102] M. A. Haq, S. J. Ruan, W. J. Cai, and L. P. H. Li, "Using lip reading recognition to predict daily mandarin conversation," in *IEEE Access*, vol. 10, pp. 53481-53489, 2022.

[103] S. Zhang, Z. Ma, K. Lu, X. Liu, J. Liu, S. Guo, A. Y. Zomaya, J. Zhang, and J. Wang, "HearMe: accurate and real-time lip reading based on commercial RFID devices" in *IEEE Transactions on Mobile Computing*, (Early access), 2022.

[104] Peng, J. Li, J. Chai, Z. Zhao, H. Zhang, and W. Tian, "Lip reading using deformable 3d convolution and channel-temporal attention," in *Proceedings of the 29th International Conference on Artificial Neural Networks and Machine Learning (ICANN 2022)*, Sofia, Bulgaria, Jul. 10-13, 2022, pp. 707-718.

[105] Xue, S. Hu, J. Xu, M. Geng, X. Liu, and H. Meng, "Bayesian neural network language modeling for speech recognition" in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2900-2917, 2022.

[106] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," *IEEE Access*, vol. 9, pp. 121184-121205, 2021.

[107] T. Ozcan and A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 2, pp. 86-90, 2019.

[108] Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning. image and vision computing," *Image and Vision Computing*, vol. 78, pp. 53-72, Nov. 2018.

[109] Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine learning (ICML '06)*, Pittsburgh, PA, USA, Jun. 25-29, 2006, pp. 369-376.

[110] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421-2424, 2006.

[111] Berkol, T. Tümer-Sivri, N. Pervan-Akman, M. Çolak, and H. Erdem, "Visual lip reading dataset in Turkish," *Data*, vol. 8, no. 1, p. 15, 2023.

[112] P. Booth, *An Introduction to Human-Computer Interaction*, Hove, UK: Lawrence Erlbaum Associates, 1989.

[113] G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech, Language, and Hearing Research*, vol. 11, no. 4, pp. 796-804, 1968.

[114] R. D. Easton and M. Basala, "Perceptual dominance during lipreading," *Perception and Psychophysics,* vol. 32, no. 6, pp. 562-570, 1982.

[115] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, Jul. 21-26, 2017, pp. 6447-6456.

[116] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui, "Audio-visual speech recognition using lip information extracted from side-face images," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2, pp. 16-21, 2007.

[117] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep learning-based automated lip-reading: A survey," *IEEE Access,* vol. 9, pp. 121184-121205, 2021.

[118] L. Pandey and A. S. Arif, "LipType: A silent speech recognizer augmented with an independent repair model," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21),* Yokohama, Japan, May. 8-13, 2021, pp. 1-19.

[119] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: end-to-end sentence-level lipreading," *arXiv preprint arXiv:*1611.01599, 2016.

[120] Shrestha and L. Rothkrantz, "Visual speech recognition automatic system for lip reading of dutch," *Journal on Information Technologies and Control*, vol. 7, no. 3, pp. 2-9, 2009.

[121] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Proceedings of the 10th International Conference on Computer Vision*, Beijing, China, 2005, pp. II: 1424-1431.

[122] M. Faisal and S. Manzoor, "Deep learning for lip reading using audio-visual information for urdu language," *arXiv preprint arXiv*:1802.05521, 2018.

[123] T. Ozcan and A. Basturk, "Lip reading using convolutional neural networks with and without pre-trained models," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 2, 2019.

[124] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Applied Sciences*, vol. 9, no. 8, p. 1599, 2019.

[125] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv*:1803.01271, 2018.

[126] Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, May. 4-8, 2020, pp. 4077-4081.

[127] Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM," Technical report, Stanford University, CS231n project report, 2016. [Online]

[128] Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML),* Pittsburgh, Pennsylvania, USA, June 25-29, 2006, pp. 369-376.

[129] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America,* vol. 120, no. 5, pp. 2421-2424, 2006.